The Molecular Landscape of HPV-driven Tumourigenesis.

Ankur Ravinarayana Chakravarthy MSc

2016

Submitted to University College London in fulfilment of requirements for the award of the degree of Doctor of Philosophy.

Declaration

I, Ankur Ravinarayana Chakravarthy, confirm that the work presented in this thesis is my own. All external sources of the information presented in this thesis have been acknowledged. Some portions of the work presented in the thesis have resulted from collaborative work and these have been acknowledged and appropriately referenced where applicable. I declare the work presented in this thesis is true to the best of my knowledge.

Ankur Ravinarayana Chakravarthy.

Abstract

Human Papillomaviruses (HPV) contribute significantly to the global cancer burden, causing nearly all cervical cancers and varying proportions of head and neck, and other anogenital cancers. Previous studies, limited to small sample sizes or single tissue sites in studies, have suggested that HPV+ tumours can exhibit distinct molecular profiles. Initially, a comprehensive transcriptional signature was established for HPV-driven tumourigenesis and culminated in the confirmation of a driver role for HPV outside the Oropharynx in Head and Neck cancer and the discovery of prognostically relevant differences in the immune microenvironment with implications for patient management. Analysis of exomes for mutagenesis by *APOBEC3B*, found upregulated in HPV+ tumours, identified it as a key driver of genomic evolution in these tumours, and broadly as a determinant of hotspot specificity in *PIK3CA* mutations across cancers.

Establishment of HPV-associated DNA methylation signatures, useful for classification, also highlighted novel HPV-related transcriptional changes and putative heterogeneity in cell-of-origin associated with HPV types and tissue sites. Analyses of transcriptional and epigenetic heterogeneity in Cervical Cancer, where multiple HPV types are causal, identified HPV45-associated molecular signatures indicative of increased invasiveness and inflammation with putative applications in patient stratification. Finally, Support Vector Regression approaches developed to perform deep deconvolution of the cellular composition of tumours facilitated integrative analysis of immune cell infiltration patterns, the prognostic patterns discovered in the thesis and molecular variation, offering insights into the role of the immune microenvironment in shaping the evolution of these tumours.

Acknowledgments

The last three years have been some of the very best I have ever had, and never have I had my intellect challenged, nurtured, and encouraged to grow as much as during this period. I am particularly indebted to these people, and non-sapient entities…

Tim Fenton, my supervisor, for always being supportive and encouraging, for being a source of wisdom, for helping to set up a space where I could sate my curiosity by following up on the research questions I found most stimulating, for encouraging me to broaden my skills above and beyond the requirements of my PhD project, for putting up with my penchant for intellectual Akathisia, and for inviting me to do a PhD in the first place. A finer mentor and role model I could not have asked for.

UCL, my intellectual home for the past few years, for being, through their Graduate and Overseas Research Scholarships, the dominant source of the funding that has enabled me to work towards a PhD. Debbie Fund, The Rosetrees Trust, Cancer Research UK and the NIHR Biomedical Research Centre for all playing critical roles in funding the research in this thesis and for keeping me safe from a state of monetary penury.

Stephen Henderson, for teaching me many of the basic skills I needed to be able to do the science I wanted and for his expertise that was essential to making significant portions of this work possible, Andrew Feber, for valuable discussions and suggestions, along with his own published work, which inspired and helped refine the science in this thesis and laid the foundations for novel-findings, and for raw data of Penile Cancer methylomes.

Xiaoping Su, for contributing vital datasets on HPV-related transcription, Matthias Lechner, the late Helga Salvesen, Martin Widschwendter and UCL Genomics for the generation of raw data/datasets that powered much of the statistical research in this thesis.

Professor Gareth Thomas, for histopathological analyses that were critical to confirming the exciting implications of transcriptional variation in immune microenvironment-related genes as a prognostic determinant in HPV+ head and neck cancers.

Alicia Oshlack, Sanja Farkas, and Professor Tatsuhiro Shibata, for generously offering access to raw data or unpublished data that contributed substantially to the power of the analyses carried out in this thesis.

Professor Kerry Chester for her supervision and for encouraging an interest in immunology, Professor Stephan Beck for encouraging and fostering a lifelong interest in Epigenetics through his module on the MSc and helping me acquire competence in computational biology through an internship after my MSc and Professor Robin Weiss for being my tertiary supervisor.

Luke Williams, the other long-term fixture in the Fenton group, for friendship, the rewarding nature of collaborative cutting-edge non-thesis science, and much mirth from a series of unfortunate accidents involving risky combinations of ambulatory contraptions and ethanol suspensions.

Professor Chris Boshoff and the members of, and friends from, the erstwhile Viral Oncology laboratory at the UCL Cancer Institute where I completed most of the first year of my PhD who were instrumental in making the lab a welcoming place, and Julie Olszewski, who has been a very helpful and kind Departmental Tutor since the days I started the Master's degree that preceded my PhD.

The many friends and comrades at the UCL Cancer Institute who have inspired me to approach research problems in novel ways, who have been in times of happiness and times of sadness, a source of support, shared happiness and shared sorrow, and critically, shared inspiration. The peer-reviewers that helped ensure rigour in my scientific work despite their tendency to occasionally be annoying.

Lacey Anderson, my beloved fiancee', for being an absolutely brilliant and loving partner, friend and confidante, who I have had the pleasure and the comfort of sharing every up and down on the PhD with, for being enthusiastic about my career plans, and for putting up with my occasionally uncontrollable workaholic tendencies.

Finally, I would like to thank my parents for encouraging me to pursue my dreams of being a scientist when so many Indian parents are besotted with Engineering and Medicine, which I personally find dreary, and for vital financial support that was instrumental to getting to a position where I could do this PhD.

Dedicated to Lacey Anderson, my beloved fiancee'

"Science is not only a disciple of reason,

but, also, one of romance and passion"

- Stephen Hawking

Table of Contents

Declaration 2

Abstract 3

Acknowledgments 4

List of Figures 18

List of Tables. 22

Abbreviations 23

Chapter 1: Introduction

1.1 Human Papillomaviruses: An introduction 29

1.2 Classification and properties of human papillomaviruses. 29

1.3 The genome structure of Human Papillomaviruses: a synopsis. 31

1.4 The Burden of HPV-induced cancers 32

1.5 Cancers are defined by a distinct set of hallmarks. 32

1.6 Cancer is a disease of genes and genomes. 33

1.7 Cell cycle checkpoints 34

1.8 High-risk HPV E6 and E7 dysregulate the cell cycle and generate the initial hits needed for malignant transformation. 34

1.9 Basic genome structure and organisation in the human genome. 36

1.10 Histone modifications and chromatin states regulate gene expression 37

1.11 An overview of DNA methylation.....39

1.12 Additional cellular changes beyond constitutive E6/E7 expression are required for progression.....42

1.13 HPV-specific pan-tissue transcriptional and epigenomic profiles - a background.44

Core Hypotheses.....45

Chapter 2: Methods

2.1 Meta-analysis to identify a pan-tissue HPV-specific transcriptional profile.....47

2.2 Validation of classification tissue independence using an expression dataset from engineered tMSCs. 47

2.3 Validation of feature selection using TCGA RNA-seq data. 49

2.4 Downstream Analysis using Ingenuity Pathway Analysis. 50

2.5 Modelling the HPV signature in the context of general malignant transformation. 50

2.6 Analysis of Methylation patterns by anatomic subsite.....51

2.7 Analysis of genomic profiles.....51

2.8 Analysis of survival associations with subsite in the TCGA cohort. 52

2.9 Analysis of immune infiltration and outcomes.....52

2.10 Exome-data preprocessing and curation for studies of APOBEC-mediated mutagenesis. 54

2.11 Quantification and examination of patterns of APOBEC mediated mutagenesis. 54

Supervised signature extraction 55

Calculating significance of enrichment using background likelihood of TCW mutagenesis 55

Generalised Linear Modelling 56

2.11 Testing whether APOBEC mediated mutagenesis is a generalised antiviral response. 56

2.12 Analysis of Driver Mutations 56

Examination of mutational trends in candidate drivers 56

Analysis of *PIK3CA* Hotspot Mutation Distributions 57

2.13 Examination of factors modulating APOBEC mediated mutagenesis 58

Viral gene transcription. 58

Subtype specific associations with overall mutational load 58

Associations with expression of APOBEC3 family genes 58

2.14 Assembly of a dataset to define HPV-associated methylation profiles. 59

2.15 Preprocessing and Normalisation of validation datasets 59

2.16 Identification of Methylation Variable Positions (MVPs) 59

2.17 Development of fDMR: an annotation-based approach to call Differentially Methylated Regions (DMRs). 61

2.18 Identification of DMR signatures using fDMR 61

2.19 Analysis of global trends in DNA methylation dysregulation. 62

2.20 Pathway Analyses....62

2.21 Machine learning....63

2.22 Independent validation of signatures....63

2.23 Analysis of distal regulatory changes....64

2.24 Integration with matched Expression Data....64

2.25 Pathway Analyses of Methylation Signatures....65

2.26 Focused Analysis of Interesting Candidate DMRs....65

2.27 Analysis of Cell-of-origin signature patterns in HPV+ Cancers....65

2.28 Assembly of cervical cancer datasets for investigating taxonomic correlates of clinical behaviour....66

2.30 Modelling taxonomic correlates of F30 status....67

2.31 Survival analyses....68

2.32 Gene expression modelling of HPV type heterogeneity....68

2.33 Methylation modelling....69

2.34 Ingenuity Pathway Analysis....69

2.35 Cohort-wide clustering analyses using HPV45-associated signature....70

2.36 Survival Analyses of signature-derived clusters....70

2.37 Development of a methylation biomarker....71

2.38 Development of a methylation signature for in-silico deconvolution....72

Dataset Assembly and Preprocessing....72

Derivation of signature features 72

2.39 Running Deconvolution Experiments using CIBERSORT 74

2.40 Estimating accuracy of MethylCIBERSORT 74

2.41 Analysis of Immune Cell Fractions based on Cervical Cancer Aggressiveness Clusters 75

2.42 Integrative Clustering and Enrichment/Overlap Analyses. 75

2.43 Omic-Signatures for Immune Clusters 76

2.44 Survival Analyses of Immune Clusters 76

Chapter 3: HPV-driven transformation is marked by a conserved pan-tissue signature of transcriptional dysregulation. 78

Chapter 3

Meta-analysis identifies a 179 feature signature of HPV-associated transcriptional changes. 79

The expression of HPV-signature genes is modulated by the expression of E6/E7 in concert with additional oncogenic hits. 81

Canonical Pathway Analysis implicates pathways associated with the known biology of HPV positive cancers. 84

Upstream regulator analysis identifies known and novel candidate upstream regulators associated with HPV oncoprotein function. 87

Epigenetic modifiers are represented in the metasignature and are putative therapeutic targets. 90

Validation of the meta-signature reveals reliable classification performance. 91

A large subset of the Metasignature is uniquely dysregulated in HPV+ tumours. 92

HPV transcript-positive OPSCC and non-OPSCC share common transcriptomic/methylomic patterns. 94

Genomic analyses further support an aetiological role for HPV outside the Oropharynx 96

HPV+ HNSC may show subsite-associated prognostic differences. 98

Anatomic subsite is associated with differences in TIL levels and activity 100

Chapter Conclusions 106

Chapter 4

Multiple genes in the gene expression signature suggest involvement of APOBEC-activity in mutagenesis and genomic evolution. 109

APOBEC mediated mutagenesis is significantly enriched in HPV+ HNSC relative to HPV- HNSC. 110

APOBEC-mediated mutagenesis is not part of a generalised antiviral response. 112

Enrichment for APOBEC-mediated mutagenesis is retained across mutational profiles of candidate driver mutations 114

APOBEC-mediated mutagenesis can determine mutational profiles and hotspot preference in cancer drivers. 116

Analysis of putative factors influencing APOBEC mediated mutagenesis.....119

Chapter Conclusions. 122

Chapter 5

Statistical analyses establish a comprehensive catalogue of DNA methylation changes in HPV-driven tumours and point towards a hypermethylator phenotype. 125

The Global Hypermethylator Phenotype Extends to Most Categories of HM450k probes. .127

Pan-tissue epigenetic signatures are useful for classification by HPV status.....129

Integrating methylation with matched expression data identifies novel HPV-associated transcriptional changes.....134

Pathway Analysis contextualises the contribution of DNA methylation to transcriptional dysregulation in HPV+ tumours. 136

A tale of two populations: Follow-up analysis on KRT7 methylation patterns suggests HPV-type associated tropism for distinct precursor cell populations. 140

Chapter Conclusions 143

Chapter 6

Molecular heterogeneity is a function of taxonomic variation in Human Papillomaviruses. 146

Pathway analysis points to an inflammatory phenotype in HPV45 driven tumours. 149

The HPV45 transcriptional signature points to metastatic behaviour 150

Upstream regulatory analysis identifies putative regulators of the HPV45 expression signature. 152

The functional context of HPV45 associated epigenomic changes. 152

HPV45 associated molecular signatures are of value beyond HPV45. 153

Machine learning yields a DNA-methylation classifier for aggressiveness associated clusters. 155

Aggressiveness Clusters Span Histology despite Cell-of-origin preferences. 156

Chapter Conclusions 158

Chapter 7

Background 160

Methylation modelling yields an accurate cellular deconvolution approach based on Support Vector Regression. 162

Variability in Immune Cell Content is associated with Cervical Cancer Aggressiveness. 165

Integrative Analysis of Immune Cell Signatures in HPV+ tumours. 167

Survival analyses of Immune Clusters suggests they comprise prognostically equivalent states of the Immune microenvironment. 174

Chapter Conclusions 176

Chapter 8: Discussion

Pan-tissue transcriptional similarities unify HPV driven cancers. 179

Successful identification of a driver role for HPV in non-Oropharyngeal HNSCs uncovers a role for immune response in mediating outcomes. 180

APOBEC-mediated mutagenesis links transcriptional and genomic evolution of HPV+ tumours. 181

A distinct, yet complex, set of epigenetic changes defines HPV-driven tumourigenesis. 182

Epigenomic and Transcriptional Analyses of Cell-of-origin signatures has implications for putative pathways to malignancies at different anatomical sites. 184

Unifying molecular profiles may improve HPV-status detection 184

Molecular heterogeneity links taxonomic variation in HPV to clinical behaviour in Cervical Cancers. 185

Analysis of Immune Microenvironment across HPV+ cancers identifies two broad immune profiles. 187

Preservation of molecular profiles has implications for therapy. 188

Final Synthesis 189

References 197

Appendices

A1: List of IPA Canonical Pathways enriched in the HPV-metasignature 208

A2: Upstream Regulatory Analysis of HPV-Associated Metasignature……..210

A3: Flowcharts describing analysis datasets and approaches used for chapters 3,5 and 6…….211

A4: Table of clinical variables and survival analyses used for comparing HPV+ OPSCC and Non-OPSCC…….212

A5: K-M curves of survival by histology and aggressiveness cluster inferred using the HPV16-like/45-like SVM classifier ……213

List of Figures

Figure 1: A Phylogeny of Human Papillomaviruses based on L1/L2 homology. 30

Figure 2: Genome structure of HPV16 31

Figure 3: Heatmapping reveals expression differences between HPV+ samples and HPV-controls across tissue and cancer types 80

Figure 4:Visualisation of metasignature genes in transformed keratinocytes.....81

Figure 5:Multiscale bootstrap analysis of HPV transcriptional signatures. 83

Figure 6:Metasignature expression in transformed mesenchymal stem cells.....84

Figure 7: Cell Cycle Progression is remodelled in HPV+ tumours.....86

Figure 8: Network of upstream regulators inferred from the metasignature 88

Figure 9:Visualisation of classifier performance in cross-validation. 91

Figure 10: Heatmap of metasignature genes in the TCGA HNSC dataset. 93

Figure 11: Visualisation of HPV-associated molecular signatures in the TCGA HNSC cohort 95

Figure 12: Distributions of HPV E6 and E7 expression by anatomic subsite 96

Figure 13: Overall Survival in the TCGA HNSC cohort, stratified by HPV status and anatomic subsite. 99

Figure 14:TIL infiltration is greater in HPV+ OPSCC. 102

Figure 15:TIL Effector expression by subsite and anatomic status. 103

Figure 16:Clustering and survival analyses based on immune profile clustering. 104

Figure 17:*APOBEC3B* expression across tumour types.....110

Figure 18: APOBEC-mediated mutagenesis is enriched by multiple measures in HPV+ HNSC. 111

Figure 19: APOBEC-mediated mutagenesis is not a generalised antiviral response.....113

Figure 20: Distributions of TCW -> TKW fractions within whole exomes and MutSig genes by category.....115

Figure 21: Illustration of PICK3CA hotspots and sequences at hotspots 117

Figure 22: Breakdown of TCW -> TKW fractions across multiple tumour types with *PIK3CA* hotspot mutations. 118

Figure 23: Analysis of APOBEC family gene expression and correlation with mutagenesis 120

Figure 24: Relationship between overall mutational load and APOBEC mutation load stratified by HPV status.....121

Figure 25: Volcano Plot showing methylation differences between HPV+ and HPV- samples 126

Figure 26:Associations between probe categories, HPV status and hypermethylation .127

Figure 27:Analysis of per-sample average methylation values by HPV status and tissue type 128

Figure 28:Heatmaps of pan-tissue epigenetic signature in the discovery set.....130

Figure 29:Heatmaps of pan-tissue epigenetic signature in validation sets.....131

Figure 30: MDS plot of signature MVPs in the discovery and Norwegian cohorts 133

Figure 31:Plots of differentially expressed genes canonically associated with MVPs 135

Figure 32: Novel DMR candidates and relationships with gene expression 138

Figure 33: Patterns of *KRT7* expression and promoter DMR methylation patterns in HPV+ tumours.....141

Figure 34: Consensus Clustering based on a squamocolumnar junction signature.....142

Figure 35: Clade associated transcriptome and methylation signatures 147

Figure 36: Clinical and molecular differences between HPV45+ and HPV16+ tumours. 148

Figure 37: The "Cell Movement" gene set activated in HPV45+ tumours. 151

Figure 38: Patterns of clustering of HPV45-linked transcriptome/methylome markers across all HPV16/45/18+ early stage CESC, and biomarker probes selected to distinguish between HPV45-like and HPV16-like clusters 154

Figure 39: Kaplan-Meier Curves of predicted HPV45-like and HPV16-like tumours within the Norwegian Cohort of Cervical Cancers.....156

Figure 40: Heatmap of SVM-based Aggressiveness Class allocation in the Norwegian Cohort. 157

Figure 41: Feature selection for MethylCIBERSORT and validation using ABSOLUTE.....163

Figure 42: MethylCIBERSORT estimates are significantly correlated with expression of marker genes.. 164

Figure 43: Estimated Immune Cell Fractions for different Immune Cell Types vary with Cervical Cancer Aggressiveness Cluster.....166

Figure 44: Pairwise correlation matrix of infiltrating cell type abundances estimated using MethylCIBERSORT in HPV+ tumours 168

Figure 45: Distributions of Immune cells vary markedly by Immune Cluster 169

Figure 46: Heatmap of 1176 genes differentially expressed between the two Immune Clusters..... 170

Figure 47:Heatmaps of cfMVP Beta-values (left) and expression of cfMVP-associated genes by immune cluster. 172

List of Tables.

Table 1: Summary of Histone modifications profiled by ENCODE and known functional associations. Table from (ENCODE 2012) 38

Table 2: Datasets used to define a pan-tissue HPV signature by meta-analysis using effect size combination. Dataset identifiers apply to EBI Array Express/ Gene Expression Omnibus. 48

Table 3: HM450k datasets used for constructing and validating an HPV-associated pan-tissue methylation signature. Discovery set components in blue text. 60

Table 4: Machine Learning Models and Tuning Parameters for developing a model to classify cervical cancers into aggressiveness clusters..... 71

Table 5: Datasets used to define cell-type specific signatures for MethylCIBERSORT deconvolution 73

Table 6: Breakdown of Hallmark genomic alterations in CDKN2A, TP53 and CCND1. HPV+ OPSCC and Non-OPSCC display similar frequencies of alteration. 97

Table 7: Table of Cox regression coefficients and P values from additive model with all infiltrating cell types with age and stage as covariates. Significant associations in blue text. Hazard Ratios are per percent increase in estimated fraction of the corresponding cell types. 175

Abbreviations

ANOVA – Analysis of Variance

BED – Browser Extensible Data

BLCA – Bladder Urothelial Carcinoma

BMIQ – Beta-Mixture Inter-Quartile

BRCA – Breast Carcinoma

CAGE (as in CAGE-seq) – Capped Analysis of Gene Expression

CESC – Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA)

CI – Confidence Interval

CIMP – CpG-Island Methylator Phenotype

CIN – Cervical Intraepithelial Neoplasia

CLL – Chronic Lymphocytic Leukaemia

CNS – Central Nervous System

COADREAD - Colorectal Adenocarcinoma

CPM – Count(s) Per Million

dB – delta-Beta

DEG – Differentially Expressed Gene/ Differentially Expressed Genes.

DMR – Differentially Methylated Region

dsDNA – double-stranded DNA

EBI – European Bioinformatics Institute

EBV – Epstein Barr Virus

eRNA – enhancer RNA

F30 – Feber-30

FANTOM – Functional Annotation Of the Mammalian Genome

FC – Fold Change

FDR – False Discovery Rate

FunNorm – Functional Normalisation (minfi)

GBM – Glioblastoma Multiforme

GLM – Generalised Linear Model

H & E / H and E – Haematoxylin and Eosin

HCC – Hepatocellular Carcinomas

HNSC – Head and Neck Squamous Cell Carcinomas

HPV – Human papillomavirus

hrHPV – High Risk HPV

HSIL – High-grade Squamous Epithelial Lesion

IARC – International Agency for Research on Cancer

IPA – Ingenuity Pathway Analysis

ISH – In-Situ Hybridisation

KICH – Kidney Chromophobe Carcinome

KIRC – Kidney Clear Cell Carcinoma

KIRP – Kidney Papillary Carcinoma

LCR – Long Control Region

LIHC – Liver Hepatocellular Carcinoma

lncRNA - long non-coding RNA

LUAD – Lung Adenocarcinoma

LUMP – Leukocytes Unmethylation for Purity

LUSC – Lung Squamous Carcinoma

MAF (file) – Mutation Annotation Format

MSC – Mesenchymal Stem Cell

MVP – Methylation Variable Position

NK-Cell – Natural Killer Cell

Non-OPSCC – Non-OroPharyngeal Squamous Cell Carcinoma

NPV – Negative Predictive Value

OPSCC – Oro-Pharyngeal Squamous Cell Carcinoma

OR – Odds Ratio

PAM (clustering) – Partitioning Around Medoids

PANCAN12 – Pan-Cancer 12.

pdCR – promoter downstream Correlated Region

PeCa – Penile Carcinoma

PIK3CA – PI3 Kinase, alpha subunit.

PPV – Positive Predictive Value

PRAD – Prosate Adenocarcinoma

RMA – Robust Multichip Average

RSEM – RNA-Seq by Expectation Maximisation

SNV – Single Nucleotide Variant

STAD – Stomach Adenocarcinoma

TCGA – Cancer Genome Atlas

THCA – Thyroid Carcinoma

TIL – Tumour Infiltrating Lymphocyte

tMSC – transformed Mesenchymal Stem Cell

TPM – Tags Per Million

Treg – Regulatory T-Cell

TSBH – Two-Step BH

UCEC – Uterine Corpus Endometrial Carcinoma

UTR – Un-Translated Region

VIN – Vulval Intraepithelial Neoplasia

VST – Variance Stabilizing Transformation

WEX – Whole Exome

WGCNA – Weighted Gene Correlation Network Analysis

WHO – World Health Organisation

WTSI – Wellcome Trust Sanger Institute

Chapter 1: Introduction

1.1 Human Papillomaviruses: An introduction

Papillomaviruses are a distinct family of evolutionarily ancient dsDNA (double stranded DNA) viruses with 189 types known in 2010 which infect a wide range of mammalian, avian and reptilian hosts, most often causing hyperproliferative structures called papillomas when they manifest visibly, while also causing latent infections and microscopic lesions which may be underestimated due to sampling biases (Bernard, Burk et al. 2010). Human papillomaviruses show a tropism for infections of epithelial tissues and often contribute to a phenotype of hyperproliferation (Orlando, Brown et al. 2013). A subset of these viruses, designated as high risk types, are capable of driving malignant transformation of infected cells and account for nearly all cases of cervical cancer (Schiffman, Castle et al. 2007). Some high-risk HPV subtypes are also responsible for driving an epidemic of head and neck squamous cell carcinomas (HNSC) (Marur, D'Souza et al. 2010). Currently, the World Health Organisation defines 12 distinct HPV types as high-risk types with sufficient evidence for a causal role in cervical cancer (CESC: Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma) , with 6 additional types putative candidates with limited evidence for carcinogenicity in humans (Bouvard, Baan et al. 2009).

1.2 Classification and properties of human papillomaviruses.

More than a 100 different human papillomavirus types have been identified to date. Those with an affinity for mucosal surfaces are classified under alpha-papillomaviruses while those with a propensity to infect cutaneous tumours are classified under beta-papillomaviruses.

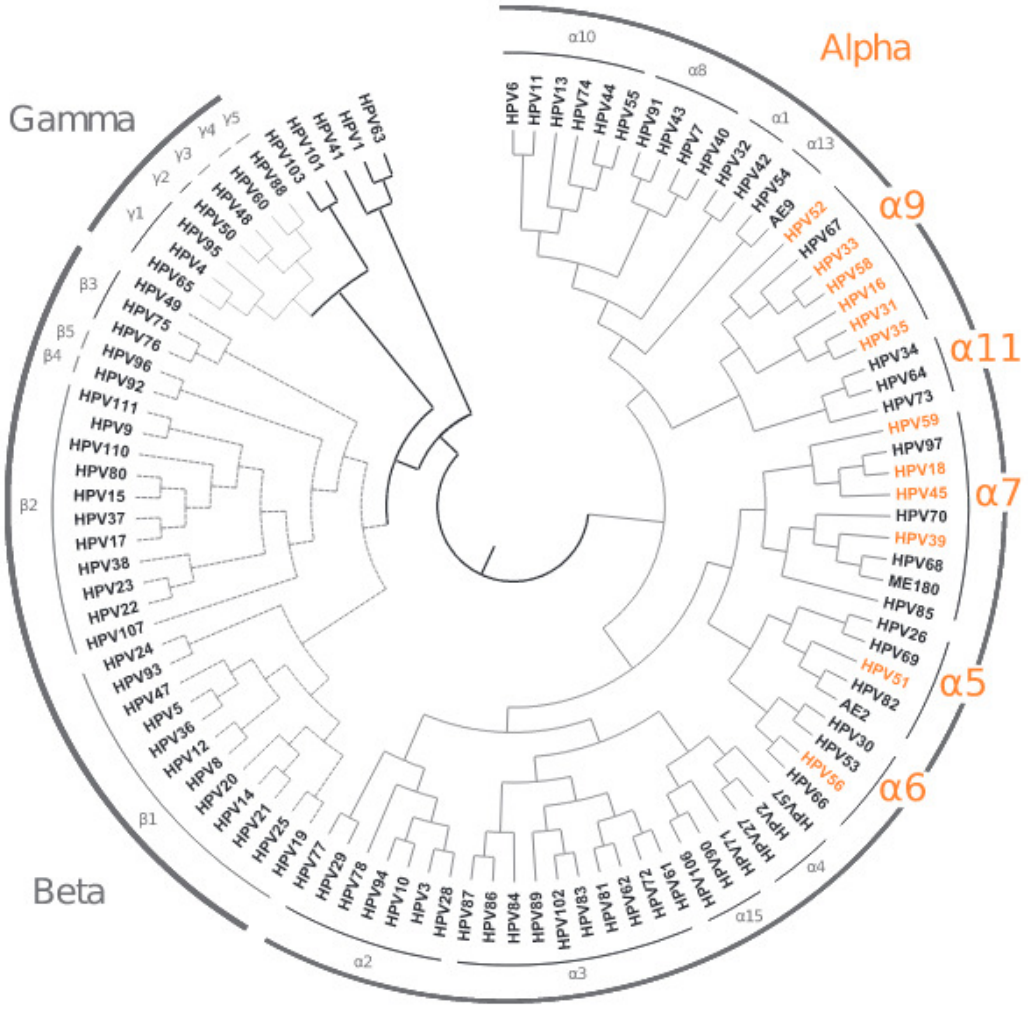


Figure 1: A Phylogeny of Human Papillomaviruses based on L1/L2 homology. Sourced from the IARC monograph on HPV. Types implicated in carcinogenesis by the WHO are labelled in orange (Bouvard, Baan et al. 2014).

Some of the alpha-papillomaviruses are associated with the development and progression of mucosal cancers and they mainly fall into "species" (In this thesis, I refer to them as Clades) Alpha-7 and Alpha-9. These types have been highlighted in orange in (**Figure 1)** and comprise "high-risk" types, the carcinogenic potential of which enjoys strong evidential support according to the WHO (Bouvard, Baan et al. 2014).

The distribution of HPV types amongst oral and anogenital tissues and cancers is subject to high levels of variability, with the exception of HPV16, which is the largest contributor to the burden of HPV-driven cancers on a pan-tissue basis (Saraiya, Unger et al. 2015).

1.3 The genome structure of Human Papillomaviruses: a synopsis.

Human papillomaviruses are dsDNA viruses with genome sizes of around 8kb. The basic organisation of the genome features two late proteins that encode the capsid (L1 and L2), a Long Control Region (LCR) and a total of eight early proteins that perform a wide range of functions (Zheng and Baker 2006) **(Figure 2)**.

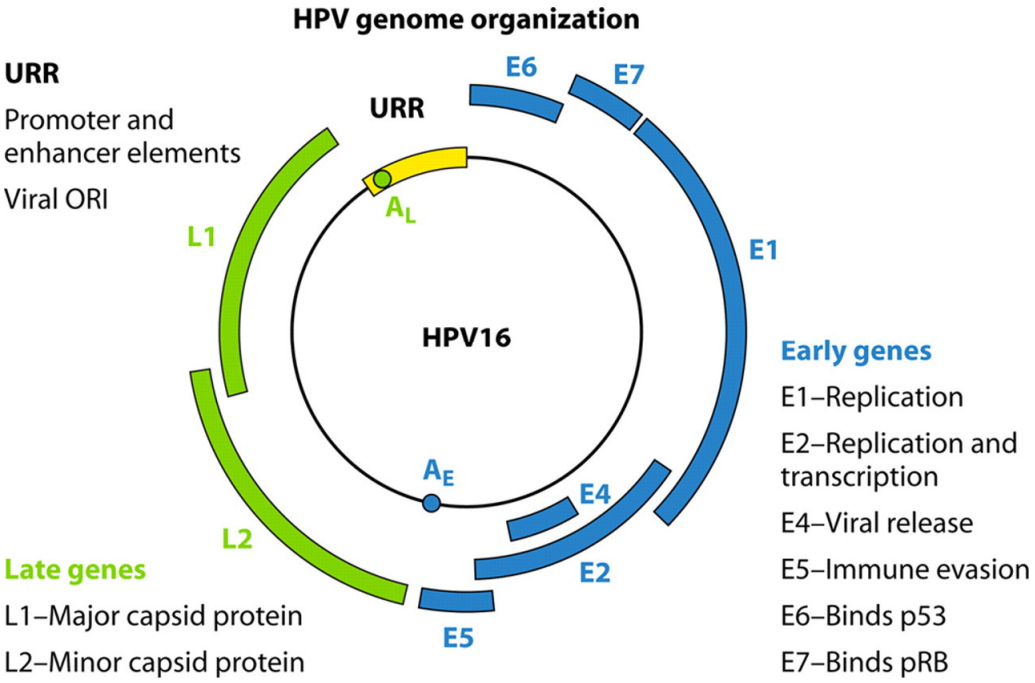


Figure 2: Genome structure of HPV16 and functions of the various proteins encoded in the genome. From (Stanley 2012)

Particularly important to HPV driven carcinogenesis are the E6 and E7 early proteins, which are described in detail in a subsequent section, and these are produced from a bicistronic RNA containing both E6 and E7. The production of E7 from this transcript requires splicing to yield a truncated E6 isoform called E6*whereas the retention of an intron is required to yield full length E6, meaning that the production of fully functional E7 can only happen at the expense of fully functional E6 (Tang, Tao et al. 2006).

1.4 The Burden of HPV-induced cancers

The burden of HPV driven cancers is the heaviest in the developing world, where access to screening may be limited, and close to 80% of CESC occurs there (Marur, D'Souza et al. 2010) with associated mortality of close to 50% (Ferlay, Soerjomataram et al. 2014). Protective vaccination is now available and has so far been demonstrated to be efficacious and capable of establishing an immune response (Paavonen, Naud et al. 2009) but is only optimally effective pre-exposure and decreased costs are an essential step towards global immunisation (Baussano, Lazzarato et al. 2013). A bivalent and a quadrivalent anti-HPV vaccine have been traditionally employed, primarily targeting, amongst hrHPV, HPV18 and HPV16 with demonstrated cross-protection against other related types, albeit one that wanes with time (Malagon, Drolet et al. 2012) and a nonavalent vaccine targeting 9 HPV types has been recently introduced (Handler, Handler et al. 2015).

1.5 Cancers are defined by a distinct set of hallmarks.

Cancer is a broad set of diseases defined by collection of common, characteristic features, often referred to as the Hallmarks of Cancer (Hanahan and Weinberg 2000).

Some of these features, including accelerated proliferation, resistance to growth-suppressive signals, evasion of apoptosis (programmed cell death), the ability to be self-sufficient in pro-proliferative signalling, creating an inflammatory microenvironment, establishing a nutritional supply through angiogenesis, replicative immortality and most importantly, to break through the basement membranes of the tissues they initially form in to then establish colonies at distant sites (metastasis) were well known in the last decade (Hanahan and Weinberg 2000). Other hallmarks such as rewiring the normal biochemistry of metabolism and evolving to evade the immune system have emerged more recently (Hanahan and Weinberg 2011).

1.6 Cancer is a disease of genes and genomes.

The hallmarks of cancer mentioned in the previous paragraph evolve by disruption of genome structure and function. These can involve DNA sequence changes, either by mutation or large structural changes that modulate how many copies of genes exist within cells. These changes can contribute to changes in the sequence, and therefore structures, of proteins encoded by these genes, thus rewiring cell signalling, or change how much of the protein is produced. Alterations in signalling networks are coupled to gene expression programs that are also known to be dysregulated in cancers. Genes dysregulated in cancer with a tangible effect on tumour fitness may either be proto-oncogenes (tumour promoting when activated) or tumour suppressors (anti-tumour activity unless disrupted). There are also genomic events that do not contribute to changes in fitness of cells that carry them in a population, which are referred to as passenger events.

Genes can also be dysregulated in cancer through epigenetic means (not involving sequence changes in the DNA of those genes) but are described separately in a later section.

1.7 Cell cycle checkpoints

The cell cycle refers to the set of all events that transpire between successive rounds of cell division. There are four distinct phases of a typical cycle of mitotic division, G1, S, G2 and M and a G0 phase where cells are not dividing. Protein signalling cascades at the transition between the G1-S checkpoint and the G2-M checkpoint are involved in regulating if a cell proceeds through the cell cycle to complete division, experience cell cycle arrest due to a wide variety of cellular and environmental conditions, or permanently exit the cell cycle. The G1-S checkpoint marks a transition between progression dependent on an external mitogenic signal to signal-independent progression (Vermeulen, Van Bockstaele et al. 2003).

1.8 High-risk HPV E6 and E7 dysregulate the cell cycle and generate the initial hits needed for malignant transformation.

The genomes of high-risk HPV strains encode multiple proteins, of which E6 and E7 are the most critical for carcinogenesis. E6 recruits the E6AP protein to form a virus specific ubiquitin ligase that targets the tumour suppressor TP53 for degradation (Kao et al, 2000). E7 binds to the retinoblastoma protein Rb and deregulates the R point transition from mitogen dependent to mitogen independent growth of the cell cycle in infected keratinocytes.

Rb, depending on phosphorylation levels, is capable of sequestering the E2F family of transcription factors, which activates a large subset of genes required to progress through to the S phase of the cell cycle. E7 is capable of directly binding to Rb via an LXCXE amino acid sequence motif and allows HPV transformed cells to bypass this cell cycle checkpoint (Harbour and Dean 2000). A key event in the development of HPV positive tumours is the integration of the HPV genome into host cells (Pett and Coleman 2007), which facilitates the loss of inhibition by the E2 protein (Collins, Constandinou-Williams et al. 2009). The E2 protein has been shown to prevent the expression of integrated HPV genomes by sequestering Brd4, a chromatin reader that recruits a transcriptional elongation factor to stimulate gene expression (Yan, Li et al. 2010) in addition to repression by direct binding to promoter elements that govern the expression of the viral early genes (Dowhanick, McBride et al. 1995). However, episomal maintenance in the cellular background of a lack of detectable integrants has been reported with HPV16 in in W12 transformed keratinocytes (Gray, Pett et al. 2010), and confirmed later in cancers, suggesting cells can still be locked into an E6/E7 expressing state without integration (Tang, Alaei-Mahabadi et al. 2013), even if they are marked by epigenetic and transcriptional differences relative to integrants (Parfenov, Pedamallu et al. 2014). To summarise, the establishment of a cellular state with constitutive E6/E7 expression leads to phenocopying the effect of genetic disruption and the subsequent loss of TP53 and pRb. This, as with other dsDNA viruses, is consistent with requirement for host cell replication machinery for viral replication, and the subsequent need to block a TP53 response, also evident in other oncogenic viruses like KSHV, Hepatitis B, EBV and SV40, that all encode TP53 degrading proteins (Munoz-Fontela, Garcia et al. 2007)

1.9 Basic genome structure and organisation in the human genome.

DNA in eukaryotes, including humans, is closely associated with nucleosomes. Each nucleosome is made of an octameric histone heteropolymer consisting of 2 copies of histones H2A and H2B, and two each of H3 and H4, and DNA that is not directly wrapped around histones is associated with linker Histone H1. Each Histone molecule has a head and outwards-facing amino acid tails with residues that are modified postranslationally, leading to varying states of compaction and relaxation. Cellular chromatin is on a more global level found folded in three dimensions, and patterns of binding of the insulator protein CTCF result in the genome being divided into domains or gene compartments and high levels of looping (Sanborn, Rao et al. 2015). On a much smaller scale, genes have upstream promoters that enable the assembly of RNA-polymerase complex and subsequent transcription of the gene downstream, which following splicing of intronic sequences and translation culminates in protein synthesis. Chromatin looping facilitates interactions between enhancers and promoters despite large linear distances (up to 500kb) separating these elements and can fine-tune gene expression(Mora, Sandve et al. 2015).

There are multiple biological processes that impact the organisation of the genome and chromatin compaction, facilitating changes in gene expression programmes. These include changes that directly modify DNA, such as DNA methylation, or histone modifications that, by altering the charge on the amino acid tails of histones, can affect compaction. Other epigenetic modifications involve deposition of novel histone variants or depletion of nucleosomes through the activity of nucleosome remodelling ATPases.

1.10 Histone modifications and chromatin states regulate gene expression

When investigating the regulatory architecture of the human genome, the ENCODE project consortium profiled multiple histone modifications **(Table 1)** that contained enough information to account for variations in gene expression to a sufficiently high degree of accuracy (Regression R ~ 0.78 in K562 cells) (ENCODE 2012) .

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements

H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

Table 1: Summary of Histone modifications profiled by ENCODE and known functional associations. Table from (ENCODE 2012)

Mutations in chromatin modifiers have been widely documented in cancer, with nearly 34 known regulatory factors dysregulated in ways consistent with a driver role across a large series of tumours from multiple sites (Gonzalez-Perez, Jene-Sanz et al. 2013). Other studies have documented that major Nucleosomal-remodelling ATPases such as members of the SWI/SNF complex are widely mutated (Kadoch, Hargreaves et al. 2013). Beyond the direct mutagenesis of chromatin modifiers, a range of pro-oncogenic dysregulation of epigenomic control is also achieved by substrate level modification, including transcription-factor binding associated super-enhancer formation through core transcriptional coactivator recruitment through intronic mutations (Mansour, Abraham et al. 2014) or changes in genome-wide histone deposition and transcriptional antigen through trapping chromatin modifiers, as evident in H3k27 mutations in paediatric gliomas (Lewis, Muller et al. 2013).

1.11 An overview of DNA methylation.

DNA methylation, involving the methylation of the 5th carbon atom of cytosine to yield 5-methylcytosine, has been one of the most well characterised epigenomic modifications. DNA methylation is induced by the activity of DNA methyltransferases, of which there are three main examples in humans; the maintenance DNA methyltransferase DNMT1 and the de-novo methyltransferases DNMT3a and DNMT3b.

DNA methylation, depending on the genomic context of the modified bases, has been shown to influence gene expression in contrasting ways. Traditionally, the hypermethylation of promoters in CpG islands (Regions of high CpG density that remain unmethylated and may be proximal to genes) has been implicated in the silencing of gene expression (Newell-Price, Clark et al. 2000), whereas the methylation of gene bodies has been associated with increased transcriptional activity (Yang, Han et al. 2014). More recent studies have highlighted that the impact of methylation on gene silencing is even more strongly evident at enhancers relative to promoters; distinct patterns of hypomethylation have been found around so-called “super-enhancers” and methylation of CTCF binding sites has been shown to result in novel, pro-oncogenic enhancer-promoter interactions (Flavahan, Drier et al. 2016).

The ability of DNA methylation to act at long distances in regulating transcription of genes that are situated several thousands of base pairs downstream, comprising promoter-downstream Correlated Regions (pdCRs), has also been documented through bisulfite sequencing based studies of tumours (Hovestadt, Jones et al. 2014) .

Additionally, DNA methylation has been sometimes shown to mark genes whose status is converse to what is canonically expected, serving as a marker for accessibility of DNA methyltransferases, and consequently transcriptional machinery, to transcribed genes and appearing in the formation of distinctive regions called DNA methylation valleys (Hovestadt, Jones et al. 2014). These discoveries, taken together, identify DNA methylation as a potent orchestrator of gene expression programmes in cellular growth and function.

DNA methylation has been implicated in pathogenesis in a wide variety of cancer and non-cancer disorders, including Rett's Syndrome (Weaving, Ellaway et al. 2005) and fragile X-syndrome (Brasa, Mueller et al. 2016). Particularly in cancer, evidence has steadily accumulated for the role of DNA methylation in modulating tumourigenesis.

De Carvalho et al (De Carvalho, Sharma et al. 2012) demonstrated using DNMT1 & DNMT3b double-knockout HCT116 cells the existence of methylation-silenced genes whose expression was cytotoxic and preserved methylation states in clinical samples of tumours, suggesting methylation events could be drivers of cancer-cell survival.

Additional studies since have shown that low doses of demethylating agents such as decitabine and aza-citidine can be used to impair the long term growth of cancer cells, through reactivating methylation-silenced tumour suppressor genes (Tsai, Li et al. 2012)and targeting gene-body methylation of oncogenes(Yang, Han et al. 2014), as well as triggering an interferon-directed immune response in putative cancer-stem-cells by reactivating the expression of endogenous retroviral elements (Roulois, Loo Yau et al. 2015).

The biological role of DNA methylation may extend beyond the regulation of levels of gene expression, for instance, it has been suggested that gene body methylation may regulate exon inclusion mediated by MeCP2 binding (Maunakea, Chepelev et al. 2013). Similarly, epigenetic resetting in neural stem-cell cultures through the induction of pluripotency demonstrated the existence of some lineage-specific methylation changes required for the successful engraftment and growth of cancer-stem-cells (Stricker, Feber et al. 2013).

DNA methylation has also been shown to play a role in the genomic evolution of cancer. Large concordant losses of methylation as tissues acquire increasing degrees of malignant transformation has been described in multiple tissue types, including hypomethylation (and methylation hypervariability) of tissue specific genes that tend to show hypervariable expression in cancer (Timp, Bravo et al. 2014).

Analysis of multiregion methylation data in prostate cancer has highlighted DNA methylation variability across individuals (Brocks, Assenov et al. 2014), and stochastic methylation changes in Chronic Lymphocytic Leukaemia (CLL) have been shown to track with survival (Landau, Clement et al. 2014). Studies in additional tumour types, such as glioblastoma multiforme (Ceccarelli, Barthel et al. 2016) , ependymomas (Mack, Witt et al. 2014) and paediatric Central Nervous System (CNS)(Sturm, Orr et al. 2016) tumours have established the presence of distinct methylation subtypes with altered implications for prognosis, such as a CpG Island Hypermethylator Phenotype (CIMP) being indicative of better-prognosis GBM and worse prognosis in ependymomas, collectively suggesting that DNA methylation patterns serve as useful features for tumour class discovery with implications for clinical practise.

1.12 Additional cellular changes beyond constitutive E6/E7 expression are required for progression.

In addition to the critical dysregulation of p53 and pRb by the viral oncoproteins E6 and E7, additional mutations, gene expression, copy number and epigenetic changes characterise HPV driven cancers. DNA sequencing studies of HPV positive HNSC carried out in the laboratory revealed a very high frequency of mutations in *PIK3CA* (mutation or amplification) and *PTEN* and *FBXW7* loss in HPV positive tumours as well as copy number changes in *SOX2* (Lechner, Frampton et al. 2013) . TCGA-led exome sequencing identified *PIKC3CA* (53%) as the predominant driver in HPV+ tumours, followed by *EP300* (14%) and *NOTCH1* (11%). Other mutation-sequencing efforts in Cervical Cancers and Head and Neck Cancers have also identified sets of recurrently mutated genes in HPV+ tumours, consistent with a driver role (Agrawal, Frederick et al. 2011, Ojesina, Lichtenstein et al. 2014). In Ojesina et al (Cervical cancers), *PIK3CA* (14%), *PTEN* (6%) and *STK11*(4%) were present in 14%, 6% and 4%, *EP300* (16%), *FBXW7* (15%), *HLA-B* (9%), *MAPK1* (8%) and *NFE2L2* (4%) were significantly recurrent in squamous carcinomas, while in adenocarcinomas, *PIK3CA* (16%), *KRAS* (8%) , *ELF3* (13%) and *CBFB* (8%) were reported, suggesting some tissue-of-origin associated heterogeneity.

A host of other studies have reported distinct transcriptional changes in HPV driven cancers, most notably a study that investigated transcriptomes in HPV+ HNSC vs HPV- HNSC and normal tissue and CESC and normal tissue (Pyeon, Newton et al. 2007). Their findings suggested that HPV induces a cascade of gene expression changes involving a large number of cell-cycle genes.

Additionally, a DNA methylation signature that could distinguish HPV positive HNSC from HPV negative HNSC has been reported, and signature probes were shown, upon multidimensional scaling, to overlap with cervical cancer methylomes (Lechner, Fenton et al. 2013). Multiple concordant lines of evidence therefore point to the importance of epigenetic processes that alter gene expression in the progression and the development of HPV driven cancers.

The identification of critical mediators of such processes that were specific to HPV positive tumours would therefore stand to serve as the basis for the development of effective targeted therapeutics and further the understanding of HPV-driven tumourigenesis.

Several proteins involved in modifying chromatin and regulating gene expression have been found to be essential for the survival of HPV positive cancer cells. Histone demethylases KDM6A and KDM6B are induced by E7 and their depletion has been shown to suppress growth in CaSki, an HPV positive cervical cancer cell line (McLaughlin-Drubin, Crum et al. 2011), this has more recently been shown to be by way of being required for p16 overexpression that prevents cell death through a yet unknown Cyclin D1 associated pathway (McLaughlin-Drubin, Park et al. 2013). Additionally studies have shown the existence of distinct cellular methylation profiles in HPV driven tumours, albeit on an individual tissue level, in HNSCs (Lechner, Fenton et al. 2013) and Penile Cancers (Feber, Arya et al. 2015).

1.13 HPV-specific pan-tissue transcriptional and epigenomic profiles - a background.

Multiple studies, including those describing datasets that I used for meta-analysis (See 2.1) have identified transcriptional changes in HPV transformed tissues relative to either normal controls or HPV- cancers. Only one of these studies however, investigated both CESC and HNSC (Pyeon, Newton et al. 2007), while two previously published meta-analyses (Buitrago-Perez, Garaulet et al. 2009, Kaczkowski, Morevati et al. 2012), as well as a pilot project I carried out for a Master's thesis, were based on the direct comparison of lists of significant differentially expressed genes from different studies, selecting genes that were associated with HPV-mediated transformation in a predetermined number of published studies as being HPV associated, and were therefore power limited by the sample numbers of the individual studies.

A recent meta-analysis used a more sophisticated statistical approach but was restricted to integrating published studies of CESC expression and copy number alterations (Mine, Shulzhenko et al. 2013) and another analysis of HPV+ samples across 12 tumour types from the TCGA PANCAN12 dataset, despite the lack of normal samples or other models of HPV+ transformation, indicated crucially for my analysis that HPV+ tumours across tissue sites were closer to each other than to HPV- samples arising in similar tissues (Tang, Alaei-Mahabadi et al. 2013). Studies on epigenetic similarities across tissue sites have only been carried out on a very small scale; (Lechner, Fenton et al. 2013) demonstrated similarities across tissue sites but this was limited by the low resolution of the Illumina 27k methylation array. (Feber, Arya et al. 2015) used a simple overlap criterion between PeCa (Penile Carcinoma) and HNSC.

Core Hypotheses

In this thesis, I test the following hypotheses

A) HPV-driven cancers share common molecular profiles independent of the tissues from which they arise. These molecular profiles are reflective of HPV biology, and their establishment is actively shaped by HPV oncoproteins (Chapters 3-5)
B) If HPV can establish tissue independent molecular profiles, it follows that variation in HPV should result in molecular differences between tumours caused by different HPV types, with potential clinical impact (Chapter 6).
C) Microenvironmental influences (including the immune microenvironment) shape molecular profiles, tumour evolution and clinical behaviours in tumours despite common aetiology (Chapters 3 and 7).

Chapter 2: Methods

2.1 Meta-analysis to identify a pan-tissue HPV-specific transcriptional profile. Microarray studies featuring HPV driven transformation with raw data available from EBI Array Express carried out on the Affymetrix hgu133plus2 or U133A platforms **(Table 2)** were identified. Datasets were rma normalised using the ***affy*** bioconductor package (Gautier, Cope et al. 2004) and reduced to the 22277 probes found on both hgu133plus2 and u133a arrays. Initial feature selection was carried out by combining effect sizes using moderated limma t-tests with the R package ***MetaMA*** (Marot, Foulley et al. 2009). I selected features with Benjamini-Hochberg adjusted P-values (the false discovery rate, referred to as "FDR" hereafter) (Benjamini) of 0.01 or less and median fold change cutoffs (median of median fold changes across each dataset) above 2.

2.2 Validation of classification tissue independence using an expression dataset from engineered tMSCs.

Raw expression data for MSCs serially transformed with the addition of *hTERT*, HPV oncogenes *E6* and *E7*, SV40 small T antigen and mutant *HRAS* were downloaded from EBI Array Express (M-EXP-563) and RMA normalised. The dataset was reduced to the 179 HPV associated probes identified by our meta-analysis. Hiearchical clustering was carried out using Euclidean distance average linkage separation of samples based on the presence of HPV oncoproteins and cellular oncogenes from E6/E7 negative tMSCs was observed. The robustness of the clusters was verified using multiscale bootstrap resampling (10,000 samples) with the ***pvclust*** R package (Suzuki and Shimodaira 2006).

Accession	Sample Description	Comparisons made	Sample numbers	Platform
E-GEOD-6791 (Pyeon, Newton et al. 2007)	Clinical samples of HNSC and normal oral epithelium, Cervical Cancer and normal cervical epithelium.	CESC vs Normal Cervical Epithelium. HPV positive HNSC vs HPV negative HNSC and Normal oral epithelium	20 CESCs, 8 Normal cervical epithelium. 26 HPV negative HNSCs, 16 HPV positive HNSCs, 14 normal oral epithelium.	Hgu133plus2
E-GEOD-15156 (Kravchenko-Balasha, Mizrachy-Schwartz et al. 2009)	Immortalised Keratinocytes. Uninfected, HPV infected early passage, late passage and Benzopyrene treated late passage cells.	Early passgage, Late passage and Benzopyrene treated late passage infected keratinocytes vs. uninfected Keratinocytes	3 each of uninfected keratinocytes, early passage , late passage and benzopyrene treated late passage keratinocytes	HG_U133A
E-GEOD-3292 (Slebos, Yi et al. 2006)	Clinical samples of HPV positive and HPV negative HNSC.	HPV positive vs. HPV negative HNSC	27 HPV negative HNSC, 9 HPV positive HNSC.	Hgu133plus2
E-GEOD-5563 (Santegoets, Seters et al. 2007)	Clinical samples of Vulval Intraepithelial Neoplasia and normal Vulval epithelium	VIN vs. Normal	10 Normal vulval epithelial samples, 9 intraepithelial neoplasia.	Hgu133plus2
E-GEOD-9750 (Scotto, Narayan et al. 2008)	Cervical Cancer cell lines, Cervical cancer clinical samples and normal cervical epithelium.	Cancer vs. normal	33 CESC clinical samples, 9 CESC cell lines, 24 normal cervical epithelium.	HG_U133A
E-GEOD-7803 (Zhai, Kuick et al. 2007)	Cervical cancer clinical samples, high grade premalignant lesions and normal cervical epithelium.	Cervical cancer/HSIL vs. Normal	10 normal cervical epithelium, 7 HSIL, 24 CESC clinical samples.	HG_U133A
E-GEOD-24089	HNSC and CESC cell lines	HPV positive vs HPV negative cell lines.	2 HPV positive, 2 HPV negative.	Hgu133plus2

Table 2: Datasets used to define a pan-tissue HPV signature by meta-analysis using effect size combination. Dataset identifiers apply to EBI Array Express/ Gene Expression Omnibus.

The process was repeated with a previously published signature defined by the manual comparison of published lists of differentially expressed genes associated with HPV-mediated transformation (Buitrago-Perez, Garaulet et al. 2009) to facilitate direct comparison.

The stability of clustering in the training datasets was estimated by constructing a misclustering index. The index was constructed by using consensus PAM (Partitioning Around Medoids) clustering to find the optimum number of clusters for each dataset via an interface written using the ***clusterCons*** package. The number of clusters with the highest robustness, estimated using the clRob function, was used to group samples and the index was defined as the percentage of HPV+ samples that were classified outside the majority HPV+ cluster.

2.3 Validation of feature selection using TCGA RNA-seq data.

Level 3 RNAseqV2 Gene level normalised data were downloaded from the TCGA data portal for 566 samples (Normal head and neck, and HNSC) and parsed into a feature by sample matrix using a custom R function. HPV status for HNSC samples was obtained from Xiaoping Su (Personal correspondence).

Data were transformed using the variance-stabilizing-transformation function in the *DESeq2* package while controlling for the covariates of HPV status and cancer/normal status (Love, Huber et al. 2014). Building and testing classifiers and further feature selection were carried out using the ***caret*** R package. Classification performance was estimated using 10 iterations of 10 fold Cross-Validation using a Random Forest model with an mtry value set to 1/3 the size of the feature-set used for classification.

Given the unbalanced nature of the dataset Cohen's Unweighted Kappa was used as the performance metric. Performance against randomly sampled features was evaluated using identical tuning and cross-validation parameters to preserve within-resample correlations and 10 sets of 157 randomly drawn genes not in our signature. The meta-signature was also compared to two previously defined signatures; one by list comparison (Davies, Wagstaff et al. 2013) and one by supervised analysis of the TCGA HNSC dataset (Tang, Alaei-Mahabadi et al. 2013) using distributions of Cohen's Kappa from 10 iterations of 10-fold cross-validation and the statistical significance of differences in distributions was estimated using Wilcoxon's Rank Sum test with Benjamini-Hochberg multiple testing correction.

2.4 Downstream Analysis using Ingenuity Pathway Analysis.

I performed a core analysis using Ingenuity Pathway Analysis (Qiagen) referring to experimentally confirmed and high-confidence predicted interactions. I identified canonical pathways, activated and deactivated predicted upstream regulators and molecular networks associated with the metasignature.

2.5 Modelling the HPV signature in the context of general malignant transformation.

Along with HNSC RNA-seq data described above, a comprehensive dataset of other tumour types with at least five normal samples from the TCGA, including Lung Squamous Cell Cancers, Lung Adenocarcinomas, Colorectal Adenocarcinomas, Breast Adenocarcinomas, Prostate Adenocarcinomas, Kidney Clear Cell and Papillary Cancers, Hepatocellular carcinomas, Bladder Adenocarcinomas and Glioblastoma Multiforme samples was established from SAGE Synapse (Omberg, Ellrott et al. 2013), comprising a

combined dataset of 6354 samples. A linear model with HPV status, Cancer/Normal status and tumour type was fit using limma-trend (Law, Chen et al. 2014) on quantile normalised log2 counts per million values of gene expression and differentially expressed genes (DEGs) were defined at a 2 fold-change and FDR < 0.01.

2.6 Analysis of Methylation patterns by anatomic subsite

Methylation data were downloaded in the form of raw IDAT files for 464 HNSC with known HPV status and were then SWAN normalised. A methylation signature was then derived exclusively from the OPSCCs to discriminate between HPV+ and HPV- tumours using a custom function based on *limma*, with significant probes defined at a delta-Beta value of 0.4 and an FDR of 0.001 or less.

A Random Forest model was fit and predictions of HPV status across the entire dataset (OPSCC and non-OPSCC) were made using the majority class vote across 10 iterations of 10-fold Cross Validation when samples where held-out. The “mtry” parameter (number of features bagged per tree) for Random Forests was set to 1/3 the size of the signature features and forests were grown with 500 trees.

2.7 Analysis of genomic profiles

Segmented copy number data from Affy SNV arrays were downloaded from the TCGA data portal (Level 3) for 545 samples, CCND1 amplifications (segment mean > 0.1) and CDKN2A deletions (segment mean < -0.1) were then called.

Somatic mutation MAF files were downloaded for exome-sequenced data (n= 528) from the TCGA data portal, and samples were checked for mutations in TP53 and CDKN2A. Proportions of these genomic alterations were then defined for combinations of HPV status (as determined for RNA-seq analyses) and anatomic subsite.

2.8 Analysis of survival associations with subsite in the TCGA cohort.

Clinical data were downloaded from the UCSC Cancer Browser. Univariate Cox regression was used to test associations between anatomic subsite (dichotomised as OPSCC/non-OPSCC) within whitelisted HPV+ tumours (Samples misclassified by both Methylation and Expression signatures with low E6/E7 expression were excluded). Subsite specific molecular profiles were analysed by searching for MVPs using the aforementioned thresholds and DEGs at a threshold of 2FC, FDR < 0.01.

2.9 Analysis of immune infiltration and outcomes

Two distinct approaches were used to estimate immune cell infiltration in HPV+ tumours. CD4 (helper T-cell) and CD8A and CD8B (Infiltrating T-lymphocyte) reads were extracted from the Variance-stabilised RNA-seq data to serve as a surrogate for infiltrating lymphocyte content.

Wilcoxon's tests were used to estimate statistical significance of difference by HPV status and anatomic subsite, with Benjamini-Hochberg correction for multiple testing. Independently, H&E digital pathology slides were evaluated on a blinded basis by Gareth Thomas, a pathologist (University of Southampton) to estimate TIL scores and morphology (basaloid/poorly differentiated vs keratinising/moderately differentiated).

Assessable samples were binned into three TIL categories (Low, Moderate and High) and associations between TIL content and anatomic subsite were tested for using the Cochran-Armitage test for trend.

The activation status of TILs was estimated based on the expression of Granzymes (Granzyme A,B,H,K and M), Perforin and IL2 using Wilcoxon's Rank Sum Test with Benjamini Hochberg correction for multiple-testing. A manually curated list of Immune Checkpoint Molecules, the aforementioned effectors, and lymphocyte markers was used to consensus cluster HPV+ tumours using Partitioning Around Medoids with 2:5 candidate clusters with the clRob function in the clusterCons R package used to select the most robust cluster.

Associations between immune clusters and anatomic subsite were tested using Fisher's Exact Test. Multivariate Cox regression was used to estimate the influence of immune transcription patterns after controlling for T-stage (dichotomised into > 3 / < 3), lymph-node status (dichotomised into > N2b / < N2b), age at diagnosis and smoking status (pack years > 10 / pack years < 10) and anatomic subsite. Smoking, T-stage and node categories were defined on the basis of previous work identifying these as prognostic factors in HPV+ OPSCC (Ward, Thirdborough et al. 2014) (Ang, Harris et al. 2010).

Additional validation was carried out by clustering HPV+ samples on the basis of gene-modules defined to be operational in Squamous Cell Carcinomas in previously published work (Ottensmeier, Perry et al. 2016). Four modules of 20 genes each (M1, M5, M6, M9) that were associated with HPV status were used. Associations with survival and anatomic subsite were tested as above.

2.10 Exome-data preprocessing and curation for studies of APOBEC-mediated mutagenesis.

Level 3 somatic mutation data in Mutation Annotation Format (MAF) were retrieved from the TCGA data matrix for 506 HNSCs and 191 CESCs, along with clinical data, which, for HNSC patients, included smoking history in pack-years and age. HPV status was allocated using the presence or absence of viral transcripts as reported in Tang et al (Tang, Alaei-Mahabadi et al. 2013). E6 and E7 RNA seq read counts were obtained by personal communication with Xiaoping Su for transcript + HNSC.

Somatic mutation data for exomes from 213 virus-associated hepatocellular carcinomas (HCC) were downloaded from the International Cancer Genome Consortium's data portal (Project - LINC-JP) and were converted to be compatible with TCGA MAF files using a custom function. MAF files for 25 EBV+ (Epstein Barr Virus) Stomach Adenocarcinomas were obtained using the TCGA data portal. Putative driver genes were identified using Firehose analysis reports from the Broad Institue Genome Data Analysis Centre using MutSig CV v2.0). Duplicate entries in .MAF files were eliminated and mutations not mapping to the standard chromosomal complement were excluded. Mutations that were not SNVs were subsequently filtered.

2.11 Quantification and examination of patterns of APOBEC mediated mutagenesis.

<u>Unsupervised signature extraction using Non-negative Matrix Factorisation.</u>

The DeconstructSigs R package (Rosenthal, McGranahan et al. 2016) was used to model mutations in trinucleotide contexts in SNVs in the TCGA dataset. Data were normalised

to trinucleotide frequency in exomes and per-sample weights were extracted. The proportions of mutations attributed to different mutational signatures was then used to obtain, based on the overall number of mutations in the sample, the fraction of mutations attributable to APOBEC-associated WTSI (Wellcome Trust Sanger Institute) signatures (Signatures 2 and 13) and not.

Supervised signature extraction

The nucleotide contexts of SNVs were extracted using a custom function written by Dr Stephen Henderson (UCL Cancer Institute, Bill Lyons Informatics Centre). The preceding and succeeding bases were retrieved from the ***BSgenome.Hsapiens.UCSC.hg19*** bioconductor package and were classified as APOBEC mutations if they occurred in the TCW->TKW context on either strand (K=T/G, W=T/A). While APOBEC-mediated mutagenesis is also known to generate TCG->TKG mutations, this shares similarities with mutations produced by the spontaneous deamination of methylcytosine and hence wasn't included in the definition of an APOBEC-mediated mutation.

Calculating significance of enrichment using background likelihood of TCW mutagenesis.

As another method of testing for associations of APOBEC-mediated mutagenesis with HPV, enrichment for TCW-TKW mutations were established relative to the background fraction of cytosines in and out of the TCW context obtained from 41 base pair windows around each SNV using Fisher's exact test. Bootstrapping was performed to define the 95% Confidence Intervals of the enrichment distributions for HPV+ and HPV- HNSCs and a Wilcoxon's test was carried out on the trimmed distributions to ascertain the significance of shifts.

Generalised Linear Modelling

Generalised Linear Models were fitted to the distributions of a) TCW->TTW mutations b) TCW->TGW mutations and c) Fraction of mutations in the unsupervised cluster that resembled the candidate APOBEC signature reported in Alexandrov et al ((Alexandrov, Nik-Zainal et al. 2013) against HPV status, age and smoking history where available using a binomial distribution function through the ***glm*** R function. Odds ratios and 95% CI estimates were retrieved using a custom R function.

2.11 Testing whether APOBEC mediated mutagenesis is a generalised antiviral response.

TCW->TKW fractions were computed using the pipeline described above and mutation contexts were visualised around cytosine sites using sequence logos. Distributions were compared using Wilcoxon's Rank Sum test.

2.12 Analysis of Driver Mutations

Examination of mutational trends in candidate drivers

To ascertain if trends seen in whole-exome data were maintained across putative driver mutations, MAF files were reduced to genes identified as significant driver mutations (q value < 0.05) using MutSig analysis as documented in the relevant TCGA Firehose reports. Supervised analysis was carried out as for the whole exome data to compute fractions of candidate APOBEC mutations. Distributions were compared between

Exome-wide estimates and MutSig-specific estimates for HNSC (HPV+ and HPV-) and CESC using Wilcoxon's Rank Sum Test.

Analysis of *PIK3CA* Hotspot Mutation Distributions

Enrichment for TCW->TKW mutations in PIK3CA in HPV+ tumours relative to HPV- tumours was tested using Fisher's Exact Test. To test the association between a skew towards *PIK3CA* Helical Hotspot mutations and increased APOBEC-activity, *PIK3CA* mutation data were downloaded from the cBio portal (Cerami, Gao et al. 2012) for cancer genomics for tumours across eight tumour types (Breast Carcinoma (BRCA), Bladder Urothelial Carcinoma (BLCA), Colorectal Adenocarcinoma (COADREAD), CESC, HNSC, Uterine Corpus Endometrial Carcinoma (UCEC) and Stomach Adenocarcinoma (STAD).

For each tumour type, the number of PIK3CA mutations mapping to either the helical and kinase hotspots were summarised and converted to proportions of all PIK3CA hotspot mutations (Helical = E542K/E545K, Kinase= H1047) found in the tumour type.

Associations between helical mutation hotspot preference and APOBEC-mediated mutagenesis was tested in two ways: binomial GLMs fitted with tumour type and mutated hotspot as variables and finally a conditional independence test using the ***coin*** R package with tumour type as a blocking factor to compare differences in TCW -> TKW mutation fractions between helical hotspot and kinase hotspot mutants.

2.13 Examination of factors modulating APOBEC mediated mutagenesis

Viral gene transcription.

Normalised read counts for E6 and E7 transcripts were obtained by personal correspondence with Xiaoping Su. Associations between APOBEC-mediated mutagenesis and viral gene expression were tested using Spearman's rank correlation between normalised read counts and TCW-mutation fraction. Correlations between E6/E7 expression and APOBEC3B expression were assessed using Spearman's rank correlation. APOBEC-family members were tested for differential expression using Wilcoxon's Rank Sum Tests. Multiple testing correction was carried out using the Benjamini-Hochberg approach.

Subtype specific associations with overall mutational load

The relationship between TCW and non-TCW mutational burden was modelled using ANOVA using the ***leaps*** R package. Models were fit with total number of TCW mutations as the response variable with non-TCW mutation total and HPV status as covariates, with both interactive and additive layouts. The adjusted R^2 value was used to select the best model.

Associations with expression of APOBEC3 family genes

Read counts were isolated for APOBEC family members from the RNA-seq matrices generated for prior analyses of gene expression and transformed into log2 counts per million. Differential expression by HPV status was estimated using Wilcoxon's Rank Sum tests and correlations with expression were tested using Spearman's Rank Correlation.

2.14 Assembly of a dataset to define HPV-associated methylation profiles.

In order to uncover patterns of DNA methylation changes that marked HPV-driven tumourigenesis, I assembled an initial dataset of 844 samples with putative associations with HPV related cancers, including HPV- tumours from the head and neck, and normal cervix and head and neck cancers from the TCGA, and normal cervix samples from another published study (**Table 3)**

Raw IDAT files were between-array- normalised using the FunNorm function in the minfi Bioconductor package (Aryee, Jaffe et al. 2014), with initial noob correction of intensities using dye-swap (Triche, Weisenberger et al. 2013). Beta values were then retrieved and corrected for probe-type differences using BMIQ normalisation (Teschendorff, Marabita et al. 2013) using 10,000 reference probes for EM fitting . 794 out of the aforementioned 844 samples had matching RNA-Seq data.

2.15 Preprocessing and Normalisation of validation datasets

I obtained raw IDAT files for a previously published comparison of HPV+ and HPV- oropharyngeal cancers, HPV+ and HPV- cell lines and Penile Carcinomas. Each of these datasets were preprocessed separately, using Functional Normalisation and BMIQ normalisation with 10000 reference probes.

2.16 Identification of Methylation Variable Positions (MVPs)

I used linear modelling using the limma package with tissue and cancer/normal status as covariates, followed by empirical bayes error estimation while correcting for mean-variance trend, to obtain a list of probes associated with HPV status.

Tissue type	Method of HPV Status Detection	Sample Breakdown	Data Source
TCGA HNSC	Viral Transcript Detection (VirusSeq, courtesy X.Su)	50 normal 68 HPV+ 434 HPV-	TCGA Data Portal
TCGA CESC	Viral Transcript Detection (Tang, Alaei-Mahabadi et al. 2013) + in-house analysis (courtesy S.Henderson)	309 cancer 3 normal	TCGA data portal
Normal Cervix Samples	Normal samples presumed HPV -	16 normal	(Farkas, Milutin-Gasperov et al. 2013)
Penile Cancers (PeCa)	P16 staining + E6 ISH	26 in total, 7 HPV+ (4 > 1 copy/ cell, 3 < 1 copy/cell)	Raw data courtesy Andrew Feber
Cervical Cancers (Norway)	RNA-seq detection of viral transcription. Normal samples presumed negative.	29 Normal, 77 tumours.	In house, for samples genomically profiled in (Ojesina, Lichtenstein et al. 2014)

Table 3: HM450k datasets used for constructing and validating an HPV-associated pan-tissue methylation signature. Discovery set components in blue text.

Adjusted P values were recalculated using a custom wrapper to the two-step BH (TSBH) procedure as implemented in the multtest package, a method shown to perform well under the presence of correlated test statistics (Benjamini, Krieger et al. 2006), to achieve control of the False Discovery Rate under multiple testing. MVPs were defined at a median beta value difference (dB) of 0.2 and an FDR of 0.001.

2.17 Development of fDMR: an annotation-based approach to call Differentially Methylated Regions (DMRs).

fDMR was developed as a derivative of the Probe Lasso approach(Butcher and Beck 2015) implemented in the ChAMP (Morris, Butcher et al. 2014)Bioconductor package. In the fDMR approach, probes are grouped into functional regions (5' UTR, Promoter, Body, IGR and 3' UTR) and functional regions that possess a minimum number of significant mvps under FDR and beta-value difference cutoffs then undergo Stouffer-Liptak p.value combination of the raw P-values using a module of code originally implemented in the Probe Lasso DMR hunter (Butcher and Beck 2015), followed by further filtering to meet FDR (BH) and beta-value difference cutoffs at the DMR level.

2.18 Identification of DMR signatures using fDMR.

DMRs were defined as promoter/5'UTR/Gene Body regions with 3 MVPs at a delta-Beta of 0.20 at an FDR of 0.01 or less, and two.sided raw P-values were used for combination. DMRs were filtered for a median DMR deltaBeta of 0.20 and a DMR FDR (Benjamini Hochberg, BH) of 0.01 or less, non-significant MVPs in the same region were included for computation of false discovery rates of candidate DMRs.

2.19 Analysis of global trends in DNA methylation dysregulation.

On a global scale, the CpG sites on the Illumina HM450k array are grouped into four sets of of CpG density-associated genomic features: CpG Islands, Open Sea probes, CpG Shelves and CpG Shores, and 7 sets of probe annotations. Probes in the HPV associated MVP/DMR signatures were tested for enrichment of hypermethylation/hypomethylation in these categories relative to HPV negative samples using a binomial test with BH FDR correction, and for overrepresentation in the signature per se using Fisher's exact test, again with BH FDR correction.

To compare overall distributions, mean methylation values were retrieved for every sample within each group (normal/HPV- tumour/HPV positive tumour) within MVPs and across all probes, and comparisons were made using Wilcoxon's Rank Sum Test.

2.20 Pathway Analyses

To evaluate the functional context of functional methylation changes, the lists of methylation-associated differentially expressed genes were passed through Ingenuity Pathway Analysis (QIAGEN) and subjected to canonical pathway analysis, upstream regulatory analysis and functional analysis using experimentally confirmed direct interactions in human tissues. Differentially expressed gene-sets were used instead of MVP/DMR associated probesets to prevent previously documented issues with the discovery of spurious associations when methylation array data are used instead.

2.21 Machine learning.

For both the overall MVP and DMR associated signatures, Random Forest models were fit through a *caret* interface to the 'randomForest' function using 10 iterations of 10-fold cross validation with the .mtry parameter set to 1/3 of the total number of features, which is default.

Class membership was estimated as the majority vote of the allocated class when samples were out-of-fold using a custom R function. Performance statistics and confusion matrices were generated using the caret R package. Kappa values, PPV (True Positives / Total Positive calls) and NPV (False Negatives / Total Negative Calls)

2.22 Independent validation of signatures

I used the predict function in the *caret* R package to estimate the performance of the model trained on the discovery set on additional datasets. Heatmaps were visualised using the aheatmap function from the NMF package with manhattan distance hierarchical clustering. These predictions were applied to Penile Cancer and Norwegian Cervical Cancers described in **(Table 3).** Analysis of batch effects in the Norwegian Cohort was compared using MDS (Multi-Dimensional Scaling) plots superimposing samples reduced to MVP-associated probes from this dataset onto the TCGA dataset. Overfitting analysis was carried out by checking if halving the size of the discovery set used for model fitting improved performance in the Norwegian dataset relative to out-of-fold estimates on the half used for model fitting, and by checking if out-of-fold estimates from training on the Norwegian dataset using features derived from the discovery set were better than those returned by models trained on the former.

2.23 Analysis of distal regulatory changes.

A BED file of known robust enhancers defined by bidirectionally transcribed loci through CAGE-seq across tissues, cell lines and primary cells, performed by the FANTOM5 consortium was downloaded from the FANTOM5 portal (Lizio, Harshbarger et al. 2015). Enhancer-associated CpG sites on the 450k array were then identified using the findOverlaps function in the GenomicRanges Bioconductor package. Expression of these enhancers in tissues representative of those in methylome analyses was tested in CAGE-seq enhancer BED files downloaded from the FANTOM5 portal for CaSki, HeLa, ME-180 and D-98 (Cervical Cancer Cell Lines), Epidermal keratinocytes and Oral keratinocytes (Primary lines), Adult cervix pool (unfractionated tissue) and HSC-3, Ca9-22, SAS, and HO-1-n-1 (Head and neck cancer) cell lines. CAGE peak start sites were summed up in the intervals overlapping the corresponding enhancers to yield TPM (Tags Per Million) estimates of eRNA expression in enhancer-associated windows for confirmation these enhancers were active in at least some of the sample tissue types in the discovery set. Each enhancer was mapped to putative candidate genes by filtering for TSS-Enhancer correlation > 0.2, FDR < 0.01 in the FANTOM5 dataset, and this was used to define enhancer-MVP pairs for integration with expression data.

2.24 Integration with matched Expression Data

Expression data (RSEM fractional count estimates) were available for 794 samples overlapping the 844 used to discover methylation signatures. Values were transformed to log2 counts per million after quantile normalization and a linear model was fit with HPV status, tissue and cancer/normal status as covariates with a mean-variance trend fit

during empirical Bayes estimation to account for the non-trivial relationship seen between mean and variance in RNA-sequencing count data. Genes were defined as differentially expressed at 2-fold change at a BH-FDR of 0.01 or less. DMR and MVP-controlled genes were then defined as epigenetically regulated if a) direction of expression change was inversely related to methylation and genes overlapped FANTOM5 enhancers or promoters or b) if non-enhancer gene-body probes were methylated and expression change was in the same direction.

2.25 Pathway Analyses of Methylation Signatures

Ingenuity Pathway Analysis was used to carry out Canonical Pathway Analysis, Diseases and Functions Analysis, Network Analysis and Upstream Regulatory Analysis with direct, experimentally confirmed interaction sets defined in Human tissues with Differentially Expressed Genes that were epigenetically regulated either by DMRs or MVPs.

2.26 Focused Analysis of Interesting Candidate DMRs

Correlations between gene expression and DNA methylation were tested out by comparing log2CPM values of RNA-seq data versus DMR-wise median methylation beta-values for matched samples using Spearman's Rank Correlation. P values were corrected for multiple testing using Benjamini-Hochberg correction.

2.27 Analysis of Cell-of-origin signature patterns in HPV+ Cancers

KRT7 DMR medians and KRT7 expression were compared between HPV+ HNSC (predominantly HPV16+), HPV16+ CESC, HPV18+ CESC and HPV45+ CESC using pairwise

Wilcoxon's Rank Sum Tests. The gene set used to define Squamocolumnar Junction cells putatively identified as the source of Cervical Cancer was obtained from supplementary data from (Herfs, Yamamoto et al. 2012). HPV+ Tumours mapping to the aforementioned groups were clustered on the basis of this signature using PAM clustering, with robustness defined using the clRob function in the ***clusterCons*** R package. Binomial tests or Fisher's exact tests were used to test for associations between Cell-of-origin signature allocations and sample group. Multiple testing correction was carried out using the Benjamini-Hochberg approach.

2.28 Assembly of cervical cancer datasets for investigating taxonomic correlates of clinical behaviour.

Level 3 RSEM upper-quartile normalised count values were downloaded from the TCGA data portal for 307 CESCs and parsed into a feature-by-sample matrix using an in-house package. Methylation beta values were extracted for this subset of samples from the discovery set established for the identification of Pan-tissue HPV methylome profiles. Clinical data were isolated from the TCGA data portal. The analysis was restricted to tumours histologically classified as Cervical Squamous Cell Carcinomas.

2.28 Identification of HPV status.

For a set of samples, HPV statuses were previously published in (Tang, Alaei-Mahabadi et al. 2013) and were used to allocate HPV type. For the remainder, RNA-seq fastq files, obtained from CGhub, were aligned using STAR to a reference index consisting of the genomes of the WHO12 high-risk types by Stephen Henderson. Samples were called

positive for a particular type if E6/E7 reads were detected, and if multiple types were present the most abundant type was assigned.

2.29 Clade allocation and filtering for analysis.

HPV Types were grouped into clades, as reported in the IARC monograph on HPV (Bouvard, Baan et al. 2014), and due to sample sizes available for each clade, analysis was carried out between two clades; with types 18, 45, 39, 70, 68, 59 comprising Clade A7 and types 67, 58, 33, 52, 31, 35, 16 comprising Clade A9.

2.30 Modelling taxonomic correlates of F30 status.

The hypothesis that there are molecular patterns and clinical differences between taxonomic groupings of HPV in cervical cancer was initially tested using a previously published 30 CpG based Nearest Shrunken Centroid classifier (Feber, Arya et al. 2015) (denoted F30 hereon) that had suggestive associations with HPV16 and better prognosis in a smaller set of TCGA CESC samples.

The extended CESC cohort was classified into F30 clusters using raw beta values following Functional Normalisation (i.e. not BMIQ-normalised to correct for probe-type bias). Binomial tests to estimate skews by clade (A7 vs A9) and types (HPV16, HPV45, HPV18) relative to F30 status against a null hypothesis of no preference for F30 status. P-values were corrected for multiple testing using the Benjamini-Hochberg method. Significant skews were defined at a false-discovery rate of 1%.

2.31 Survival analyses.

For purposes of modelling, clinical stages were aggregated into distinct groups (for instance, Stage IA and IB were grouped into Stage I). For selection of potential stratifiers or covariates for inclusion in Cox models, we first used Cox regression to estimate the strength of associations between age of diagnosis, year of diagnosis, grade and aggregated stage.

Aggregated stage was the only significant covariate and was included as a covariate/stratifier in survival analyses. Stratified, additive and interactive Cox proportional models and where applicable, checked the model of no interaction was justified using ANOVA, for F30 status, Clade and Type as predictors (one at a time with aggregate stage as a covariate). The analysis was also repeated by restricting tumours to early stage (Stages I and II). The appropriateness of proportional hazards assumptions was tested using the cox.zph function.

2.32 Gene expression modelling of HPV type heterogeneity.

Cox regression modelling indicated significant differences in survival between HPV45 and HPV16, especially amongst early stage samples, with HPV18+ tumours comprising an intermediate group. To explore the extent and significance of molecular heterogeneity associated with HPV type comparisons were therefore performed between HPV16 and HPV45 + tumours.

RNA seq RSEM fractional counts were quantile normalised and filtered to remove low-expressed genes (less than 1 count per million in more than a quarter of samples).

Limma voom was then used to estimate mean-variance precision weights for robust linear modelling. Differentially expressed genes were identified at a mean fold-change of 2 and a BH-corrected false-discovery rate of 0.01.

2.33 Methylation modelling

Differentially methylated probes were identified using a custom R wrapper to limma with correction for mean-variance trend upon empirical bayes modelling to account for the heteroscedasticity of beta-values. Significant MVPs were defined at a mean delta-beta of 0.1 at a two-step BH adjusted false discovery rate of 0.01. Small beta-value thresholds were used to permit more sensitive detection of methylation differences.

The identification of DMRs was carried out using fDMR, and DMRs were required to have at least 3 mvp's at a false discovery rate of 5% and an average deltaBeta difference of 0.1 and a DMR FDR of 0.05. Both these signatures were reduced to features associated with differential expression at an FDR of 0.01 and a fold change of 2, following limma-trend (Law, Chen et al. 2014) analysis of log2-cpm matrices reduced to genes in the signatures.

2.34 Ingenuity Pathway Analysis

A core analysis was carried out across the expression of genes from the whole-transcriptome, DMR-associated and mvp-associated signatures by limiting interactions to experimentally validated ones. Canonical pathways analysis was adjusted for multiple testing during pathway selection.

Upstream regulatory networks were built using molecules with inferred states of activation/repression while controlling the p.value of overlaps for false discovery rate using the Benjamini-Hochberg method. False discovery rate correction was also carried out for any enrichment analyses involving multiple comparisons using custom R code written to analyse output of IPA analyses.

2.35 Cohort-wide clustering analyses using HPV45-associated signature

Joint clustering of the HPV45-16 gene expression signature and the functional MVP signature was carried out using a random forest proximity matrix to serve as the distance matrix across the dataset of stage I and stage II tumours. PAM Consensus Clustering was then carried out with 2:4 candidate clusters to identify the most robust partitioning of samples using a custom R wrapper calling clustering and robustness evaluation functions from the ***clusterCons*** R package.

2.36 Survival Analyses of signature-derived clusters

Cox proportional hazards regression was carried out for early stage samples across the cohort of HPV16+, HPV18+ and HPV45+ in two conditions; Firstly, all HPV45-like tumours versus all HPV16-like tumours, and secondly, all HPV45-like and HPV16-like tumours with HPV45s excluded in order to estimate utility beyond HPV45+ tumours. Independent contribution from these signatures was quantified by fitting Cox regression models with HPV type as a covariate in addition to aggregated clinical stage.

2.37 Development of a methylation biomarker.

Linear-modelling was carried out to identify probes differentially methylated between HPV16-like and HPV45-like tumour clusters at an FDR of 0.01 (Two-step BH corrected), dB > 0.3. Significant MVPs were then used to train an assortment of classifiers with the following tuning parameters **(Table 4)** using the ***caret*** R package using 10 iterations of 10 fold Cross-Validation.

Model	Parameters and Ranges
K Nearest Neighbours (kNN)	K (number of nearest neighbours) – 1,3,5
Gradient Boosted Machine (gbm)	Shrinkage – 0.01,0.1,0.5,1 ; Minimum observations in node = 10, interaction depth – 5,10 , number of trees – 150,300
Nearest Shrunken Centroid	Threshold – 0.01,0.1,1,3,5,10,20.
Glmnet (Elastic Net)	Alpha – 0,0.2,0.4,0.6,0.8,1 ; Lambda – 0, 0.01, 0.02, 0.03, 0.04, 0.05
SVM (Support Vector Machine) – Linear Kernel (svmLinear)	C – 0 to 1, intervals of 0.1.
Random Forest	mtry = 1/3 the number of features.

Table 4: Machine Learning Models and Tuning Parameters for developing a model to classify cervical cancers into aggressiveness clusters.

For each model, the best parameters were selected on the basis of maximum Kappa values from the cross-validation tuning/fitting process. Class allocation for the dataset was made using out-of-fold calls to generate a majority vote. The different models were compared on the basis of Kappa values and how closely they recapitulated the Hazard Ratios of Cox regressions carried out on type-associated clusters to select a model for application to validation datasets.

2.38 Development of a methylation signature for in-silico deconvolution.

Dataset Assembly and Preprocessing

Raw data were obtained in the form of IDAT files from a collection of sources that are summarised in **(Table 5)**.CD4+ cells were removed from the Blood.450k dataset and CD4+ T-cells from the Chen dataset were binned into categories based on whether they were Tregs (T-regulatory cells) or not. Neutrophils are a subset of Granulocytes and therefore they were aggregated into a single category for further analysis.

The files were parsed into R using the *minfi* Bioconductor package and were between-array normalised using Functional Normalisation as implemented in *minfi*. Probe-type bias was corrected using BMIQ Normalisation with 10,000 reference probes and type2 probes were also adjusted using Expectation-Maximisation fits.

Derivation of signature features

A custom limma based wrapper function was used to fit a series of linear models for one versus all other samples for each cell type. Features from this set of analyses were then restricted to MVPs that showed a median beta-value difference of 0.3 at an FDR of 0.01

for that fit or less. An additional filter was then applied based on whether probes associated with a cell type were hypermethylated or hypomethylated by a median beta-value difference of at least 0.1 compared to the closest cell type in terms of median value for that probe.

Cell Types	Data Source	Publication Reference/ Accession IDs
Granulocytes (12), CD8+ (6), CD19 (6), CD56 (6), CD14 (6), Eosinophils (6)	Blood.450k Bioconductor package.	(Reinius, Acevedo et al. 2012)
Fibroblasts (4)	Gene Expression Omnibus	GSE74977
HNSC Cell Lines (12) , HeLa	HNSC Cell Line Data from Matthias Lechner. HeLa from ENCODE	HNSC Cell Lines: (Lechner, Fenton et al. 2013)HeLa : GSM999337
CD4+ helper cells (6) and CD4+ Tregs (4)	Raw data obtained through personal communication with Alicia Oshlack	(Chen, Miao et al. 2015)

Table 5: Datasets used to define cell-type specific signatures for MethylCIBERSORT deconvolution

Features were then ordered by t-statistics and the top and bottom 50 features were picked to try and ensure a balanced sampling of hypermethylated and hypomethylated probes. Finally, for use with CIBERSORT, data were transformed from beta values (bound between 0 and 1) to percentages (0 – 100). Type-wise means were estimated for each probe and cell type and the matrix was exported for upload to CIBERSORT.

2.39 Running Deconvolution Experiments using CIBERSORT

The 844 methylome dataset, and the signature matrix derived in the previous step, were uploaded to CIBERSORT at https://cibersort.stanford.edu . The data were quantile normalised and CIBERSORT was run using 1000 permutations. Output files were downloaded as tab-delimited text files and custom parsers were used to import results into R for downstream analysis.

2.40 Estimating accuracy of MethylCIBERSORT

In the absence of flow-cytometry based estimates for the different cell types in the analysed tumours, the estimated fraction of Cancer Cells from MethylCIBERSORT was compared to sequencing-data based estimates from ABSOLUTE available for 466 HNSCs from previously published work (Aran, Sirota et al. 2015) using Spearman's Rank Correlation. Additional correlations were investigated between ABSOLUTE and other methods of estimating purity/immune cell fraction in this subset of tumours – including LUMP, ESTIMATE and H&E staining assessment. Spearman's Rank Correlation was used to estimate expression between marker transcripts (*CD19* for B-cells, *FOXP3* for Tregs, *CD8* for CD8+ lymphocytes). Where applicable, multiple testing correction was performed using the Benjamini Hochberg approach.

2.41 Analysis of Immune Cell Fractions based on Cervical Cancer Aggressiveness Clusters

The SVM trained on Cervical Cancers to classify samples into HPV45-like and HPV16-like samples was applied to the entirety of the 844 sample discovery set used for defining the Pan-tissue methylome signature for HPV-driven tumourigenesis using the "predict" function from the ***caret*** package. Immune cell fraction distributions were compared between the two allocated classes using Wilcoxon's Rank Sum Tests. Multiple testing correction was carried out using Benjamini-Hochberg adjustment.

2.42 Integrative Clustering and Enrichment/Overlap Analyses.

Feature by Sample matrices were generated for 345 HPV+ samples with methylation data available consisting of Immune Cell Fraction estimates for seven immune cell types and stromal fibroblasts, and sets of immune effector molecules and immune checkpoints as defined in section 2.9. A Random Forest was then used to jointly estimate a Sample proximity matrix that was transformed into a dissimilarity measure using square root of 1-Proximity, and was divided into clusters using PAM clustering through a custom wrapper calling functions from the ***clusterCons*** package, as described earlier, with the most robust number of clusters from the range of two to five clusters picked as estimated using the clRob function. Overlap analyses were carried out using Fisher's Exact Test.

2.43 Omic-Signatures for Immune Clusters

Limma-trend was used to model differential expression between Immune-related clusters, and DEGs were called at 2FC, FDR < 0.01. Core Pathway Analysis was carried out using Ingenuity Pathway Analysis with experimentally confirmed direct interactions in human tissues/cells serving as the reference point. MVPs were defined between the two clusters at a median delta-Beta of 0.15 and a TSBH FDR < 0.01 and were then reduced to cfMVPs based on concordant overlaps with DEGs.

RPPA protein data were obtained for 162 samples for 194 proteins from the UCSC Cancer Browser, quantile normalised, and then differential antibody staining was estimated using limma-trend as for expression, with a cutoff of FDR < 0.05.

2.44 Survival Analyses of Immune Clusters

Clinical data were downloaded from the UCSC Cancer Browser for 337 of the 356 HPV+ samples for which immune cluster, aggressiveness cluster and cellular infiltration estimates were available. Cox proportional hazards regression was used to estimate the impact of immune cell fractions (Percentages from MethylCIBERSORT analysis) on survival. The following models were evaluated…

A) Multivariate Cox regression to estimate impact of Immune cluster, with strata for stage and age at diagnosis as covariates.

B) Cox regression models fit separately in Head and Neck Cancer and Cervical Cancer with age at diagnosis and stage as covariates and all infiltrating cell type estimates set as predictors. HR estimates were derived per percent increase in cell fraction.

C) A threshold based model where estimates were split into four quartiles and comparisons were carried out between upper and lower quartiles, with age and stage for covariates. Models were fit separately for Head and Neck and Cervix.

Chapter 3: HPV-driven transformation is marked by a conserved pan-tissue signature of transcriptional dysregulation.

I published many of the findings in this chapter in the Journal of Clinical Oncology, with the paper titled “Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx.” (Chakravarthy, Henderson et al. 2016). The main analytical steps and the relationship between datasets are described in a flowchart in Appendix A3. The clinical data used for comparing HPV+ OPSCC and Non-OPSCC are summarised in Appendix A4.

Meta-analysis identifies a 179 feature signature of HPV-associated transcriptional changes.

I conducted a meta-analysis of 7 gene expression microarray studies that included tissue samples and cell line models representing normal epithelium (69), HPV- (55) and HPV+ (138) tumours **(Table 2: See Methods, page 48).** 179 probes, representing 159 unique genes, were differentially expressed in the HPV+ samples according to stringent criteria for statistical significance (FDR <0.01) and fold change in expression (FC >2) (hereby, this gene set is referred to as the metasignature). This was done to estimate the significance of genes to derive a pan-tissue HPV-associated signature while reducing the influence of platform and study-specific effects that could confound joint analysis.

This gene signature showed a consistent pattern of differential expression between HPV+ tumours and HPV- tumours or normal tissue across the clinical sample datasets used in the study as visualized using heatmaps **(Figure 3),** exhibiting general clustering. Visualisation of signature gene expression in a published dataset consisting of keratinocytes transformed with HPV16 (Kravchenko-Balasha, Mizrachy-Schwartz et al. 2009) revealed stronger expression in late passage and benzopyrene treated late-passage transformed keratinocytes than that of cells profiled immediately following HPV infection, resulting in accurate clustering **(Figure 4)**.

These data suggest that the expression signature reflects the action of HPV in concert with cellular changes that occur during transformation, rather than simply a set of genes directly modulated by HPV.

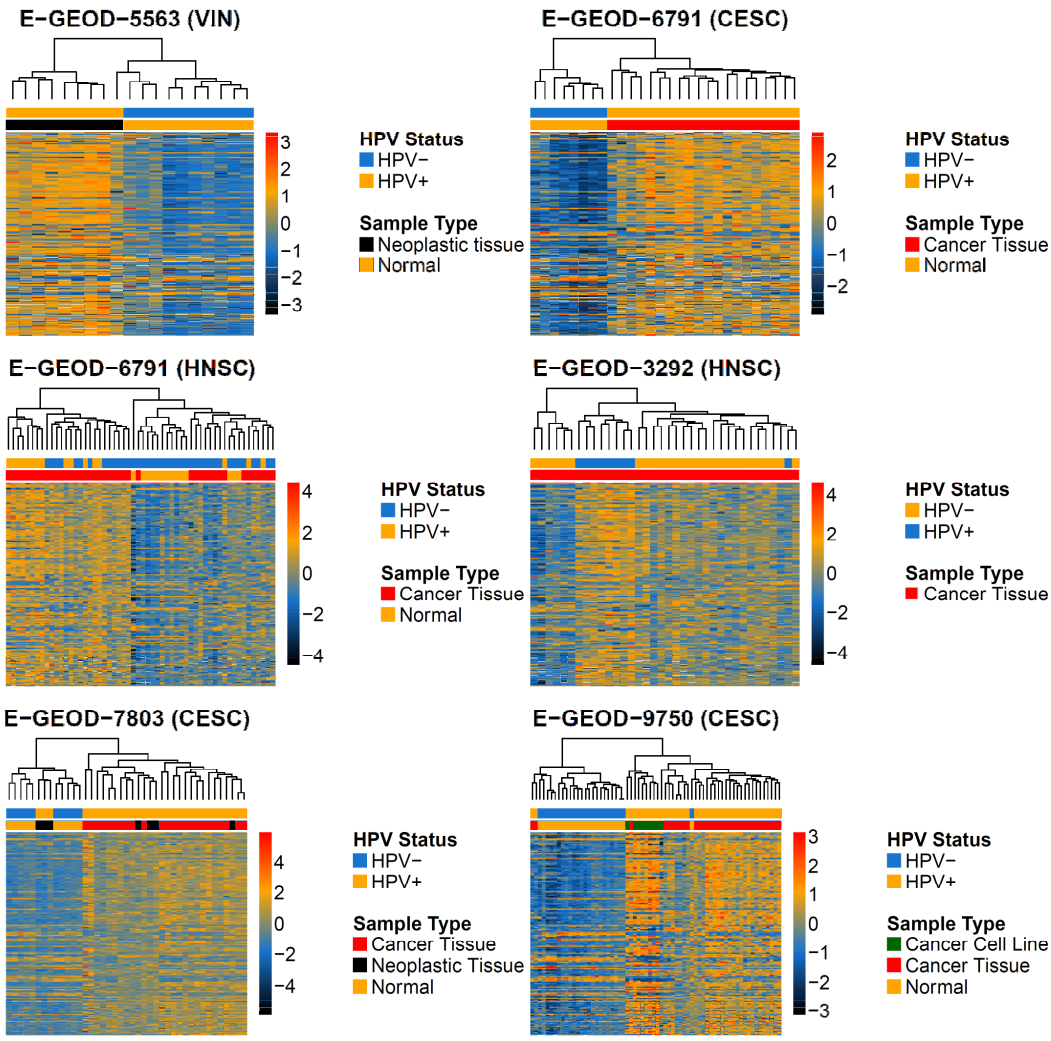


Figure 3: Heatmapping reveals expression differences between HPV+ samples and HPV- controls across tissue and cancer types. Clustering generally occurs by HPV status. In annotation ribbons, top row represents HPV status, bottom row represents sample type.

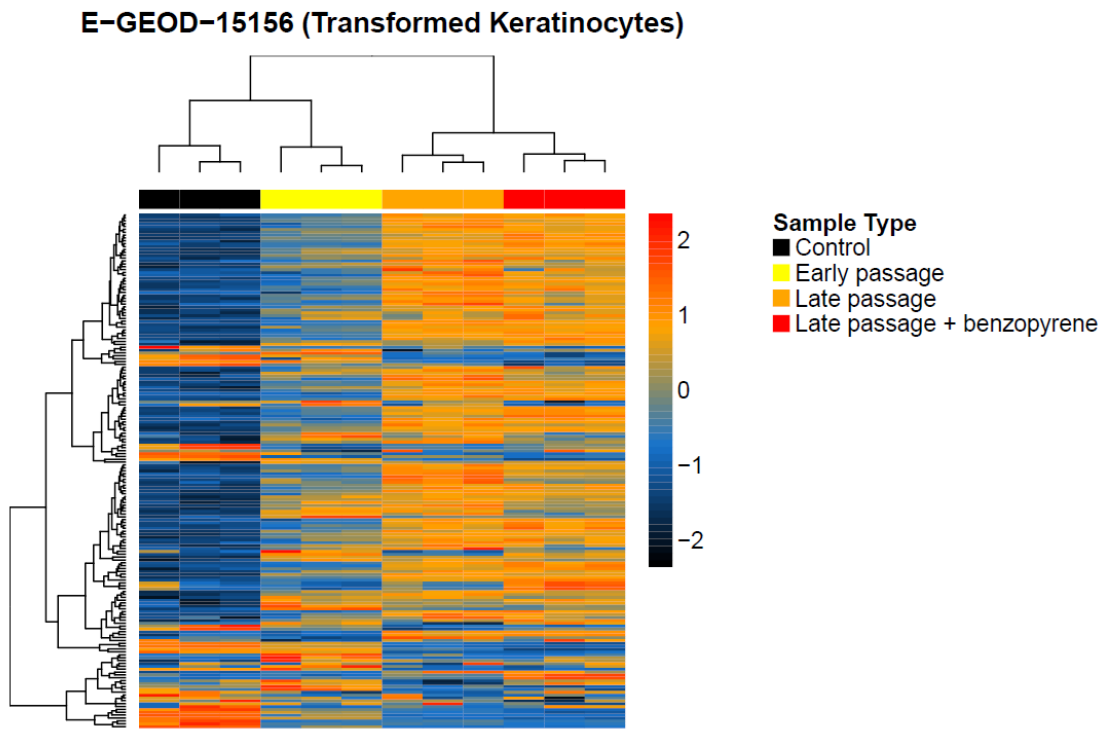


Figure 4: Meta-Signature genes show maximal expression in late-passage and benzopyrene-treated late passage foreskin keratinocytes transformed with HPV relative to HPV- controls and early passage HPV+ cells.

The expression of HPV-signature genes is modulated by the expression of E6/E7 in concert with additional oncogenic hits.

To further investigate the influence of HPV oncogene expression and additional alterations in cellular genes on expression of the signature analysis was carried out on an expression dataset derived from the stepwise transformation of mesenchymal stem cells (MSCs): unmodified MSCs; MSCs+hTert; MSCs+hTERT+E6/E7; MSCs+hTERT+E6/E7+SV40 small T, MSCs+hTERT+HPV16 E6/E7+SV40 small T+HRAS V12 (Funes, Quintero et al. 2007). Cluster multiscale bootstrap analysis (**Figure 5A**) showed the presence of three robust clusters based on expression of the HPV signature.

These clusters corresponded to: (1) primary and hTERT-immortalized MSCs; (2) immortalized MSCs expressing HPV16 E7 (with or without E6) and (3) E6/E7 expressing MSCs expressing additional oncogenes (SV40 small T and mutant HRAS). The expression of the HPV-associated signature is most closely recapitulated in cluster 3 **(Figure 6)**, i.e. cells that are expressing not only the HPV oncogenes but also the SV40 small T antigen (with or without the addition of oncogenic *HRAS*). SV40 small T inhibits the PP2A phosphatase and its primary target in transformation appears to be the stabilization of c-MYC (Yeh, Cunningham et al. 2004).

These findings from both the keratinocyte and MSC transformation models suggest not surprisingly, that the megasignature for HPV-driven tumours results from the combined action of the HPV oncogenes and the activation of cellular proto-oncogenes that are known to be important for full transformation. Additionally, the fact that the signature is recapitulated upon transformation of mesenchymal stem cells demonstrates that many of these genes are modulated during HPV-driven transformation, independent not only of the anatomical site or tissue of origin but also of the cellular lineage involved. Indeed, the intensity of signature gene expression also increases with degree of transformation in E-GEOD-7803 from CIN3 to cancer (Figure 3).

The presence of a distinct robust cluster consisting of E6/E7 positive tMSCs with additional oncogenic hits with the metasignature that clusters away from E6/E7 expressing cells alone and tMSCs but not upon clustering using features in a signature previously proposed by (Buitrago-Perez, Garaulet et al. 2009) (**Figure 5B**) suggests that the metasignature is a better approximation of full transformation in the presence of HPV than just E6/E7 expression.

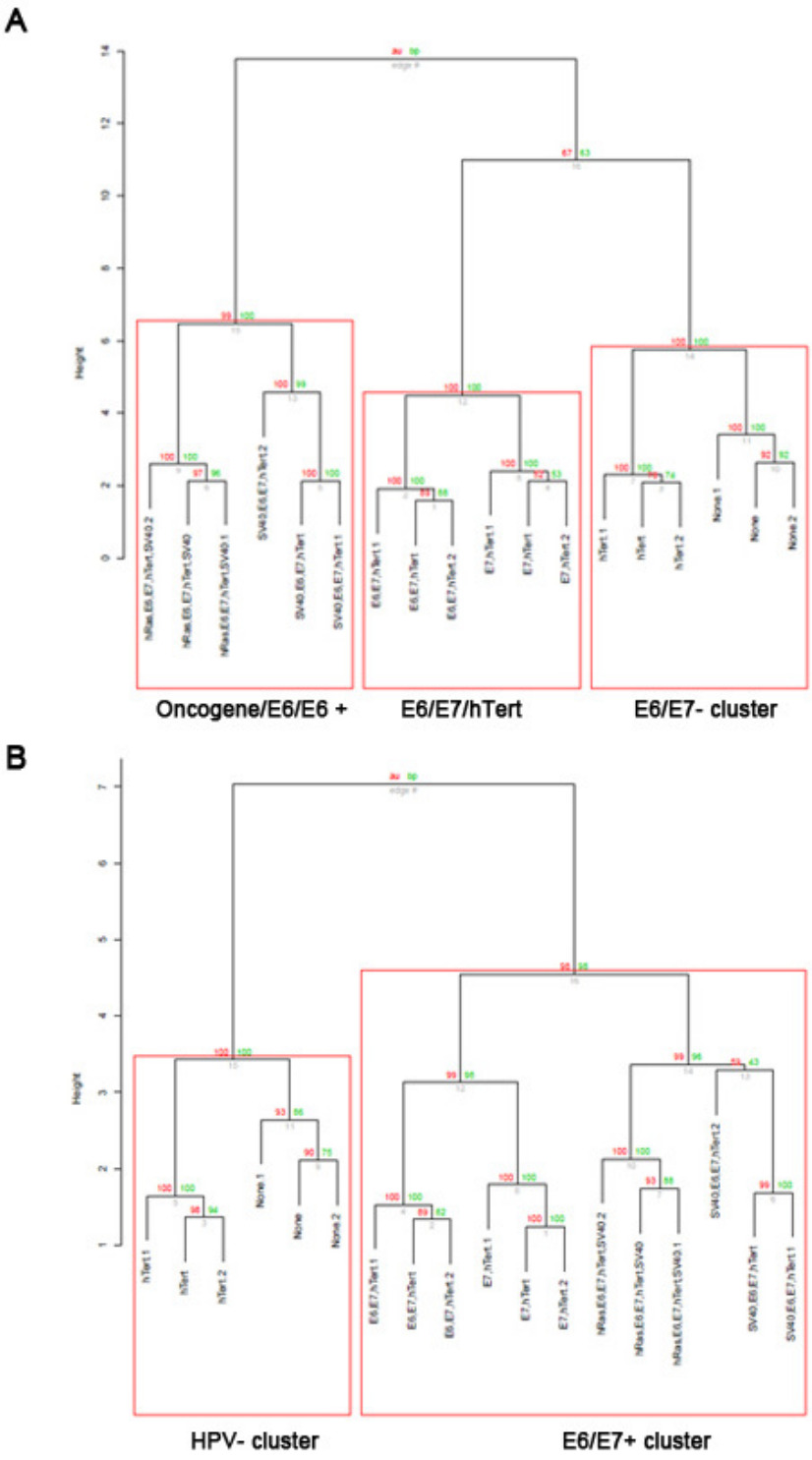


Figure 5: A) Application of the meta-signature results in stratification of tMSCs into three robust clusters (top) by average linkage euclidean distance resampling whereas a previously defined signature (B) does not (bottom). Red boxes define clusters with bootstrap stability > 0.95. E6/E7 + oncogene transformed cells represent fully transformed cells.

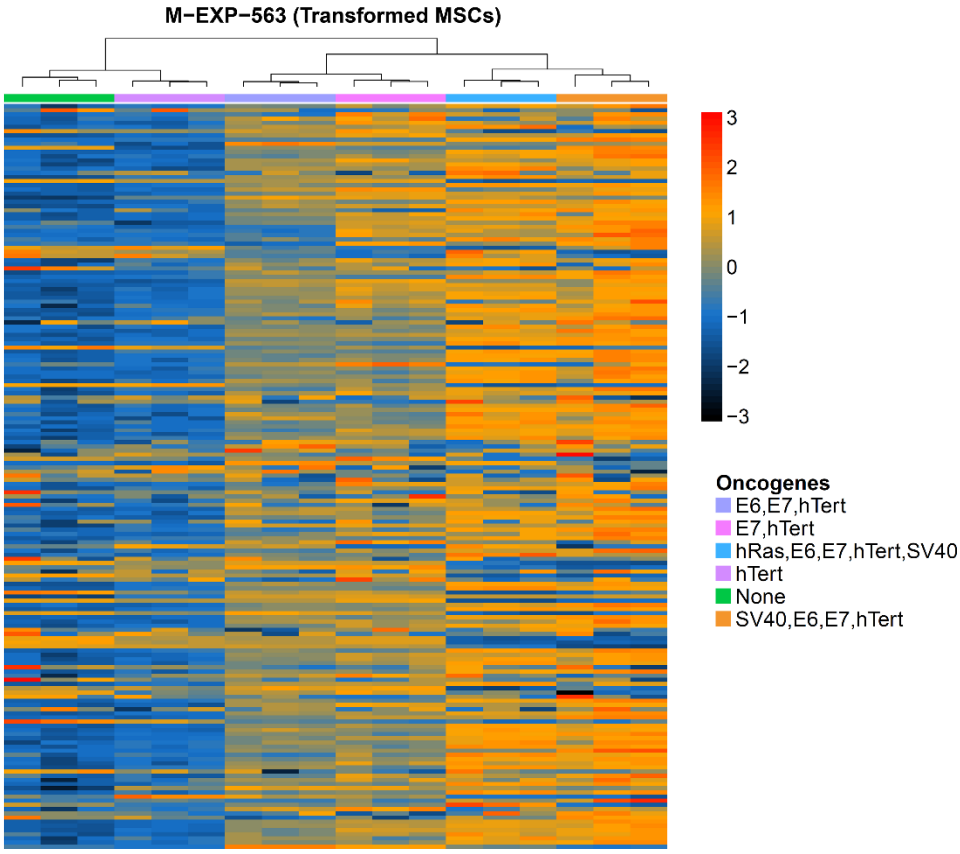


Figure 6:The HPV meta-signature is best recapitulated in cells containing E6/E7 and additional oncogenes. Intensities represent row z-scores

Canonical Pathway Analysis implicates pathways associated with the known biology of HPV positive cancers.

By deriving a list of genes whose expression becomes deregulated in HPV-driven cancers, irrespective of the anatomical site at which they occur, I aimed to identify a core set of pathways that are important for the development and maintenance of these tumours.

To determine the major pathways represented and enriched in the HPV-expression signature I performed canonical pathway analysis using the Ingenuity Pathway Analysis (IPA) software (see methods). Canonical pathway analysis of the signature genes identified multiple pathways associated with cell cycle progression, as previously noted by (Pyeon, Newton et al. 2007, Buitrago-Perez, Garaulet et al. 2009),DNA damage and DNA replication as significantly enriched amongst signature genes **(Appendix A1)**.

The mechanisms by which HPV promotes cell cycle entry and DNA replication have been intensively studied, lending to confidence in the procedure used to derive the signature by relating a number of signature genes and pathways to well-established facets of HPV biology.

Firstly, the abrogation of pRb function by E7 overcomes the need for CyclinD1-Cdk4/6 activity to drive passage through the G1/S transition and commitment to DNA replication. Consistent with this, the metasignature is characterised by reduced *CCND1* expression and upregulation of *CCNE2* & *CDK1*, which act at later stages of the cell cycle (**Figure 7**).

Consistent with HPV biology is the status of *CDKN2A* as a signature gene. While *CDKN2A* is frequently lost or mutated in HPV-negative tumours, it is induced in cells expressing E7 and the protein product $p16^{INK4A}$ is used as a biomarker for HPV+ tumours (Klaes, Friedrich et al. 2001). *CDKN2A*, in addition to being a biomarker for HPV-transformed cells, also appears to be necessary for their survival through its role in suppressing an as-yet uncharacterized CDK4/6-dependent cell death pathway (McLaughlin-Drubin, Park et al. 2013).

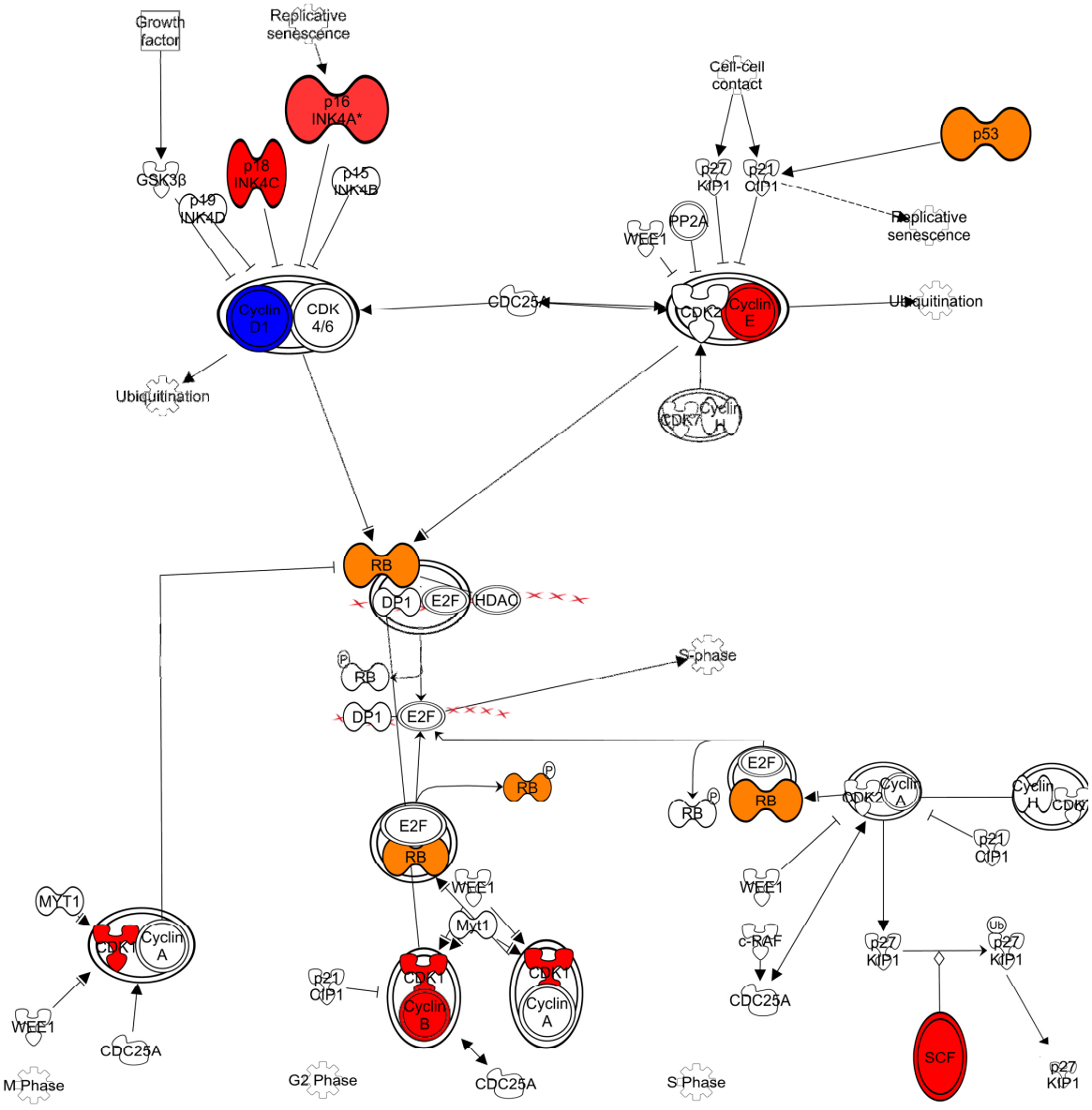


Figure 7: Cell Cycle Progression is remodelled in HPV+ tumours, with progression mainly mediated by CCNE-CDK1 complexes as opposed to CCND1-CDK4/6 complexes, which is reflected in the relative downregulation of *CCND1*. The tumour suppressor CDKN2A, *SCF*, and *CCNB1* are also strongly upregulated. Overexpressed genes are in red, underexpressed genes are in blue, and HPV oncoprotein targets (E6 and E7) are coloured in orange.

Multiple genes required for DNA replication and S-phase progression are represented in the signature, including all six MCM complex members (*MCM2-7*), *TIMELESS*, the replication factor C subunits *RFC4* and *5* and *RPA2*. The upregulation of these genes may reflect the replication stress caused caused by aberrant S-phase entry in cells expressing HPV oncogenes. The single stranded (ss)DNA-binding protein, RPA2 and another signature gene product, FANCI have both been implicated in promoting DNA repair at stalled or collapsed replication forks, a process impacted by E6 (Smogorzewska, Matsuoka et al. 2007, Day and Vaziri 2009, Shi, Feng et al. 2010) as well as *CCNE* overexpression (Ekholm-Reed, Mendez et al. 2004). Furthermore, FANCI is a component of the Fanconi Anaemia complex, which is activated in response to DNA damage caused by E7 expression (Spardy, Duensing et al. 2007, Park, Shin et al. 2013).

TIMELESS plays a key role in relaying the replication stress checkpoint signal from ATR to Chk1 while RFC4 and 5 play dual roles in DNA replication and repair, including loading of the 9-1-1 complex onto DNA in response to replication stress (Unsal-Kacmaz, Mullen et al. 2005, Mailand, Gibbs-Seymour et al. 2013).

Upstream regulator analysis identifies known and novel candidate upstream regulators associated with HPV oncoprotein function.

In addition to investigating the biological pathways in which metasignature genes participate, I also used IPA upstream regulator analysis to identify the major pathways whose modulation in HPV+ cancers induces these expression patterns.

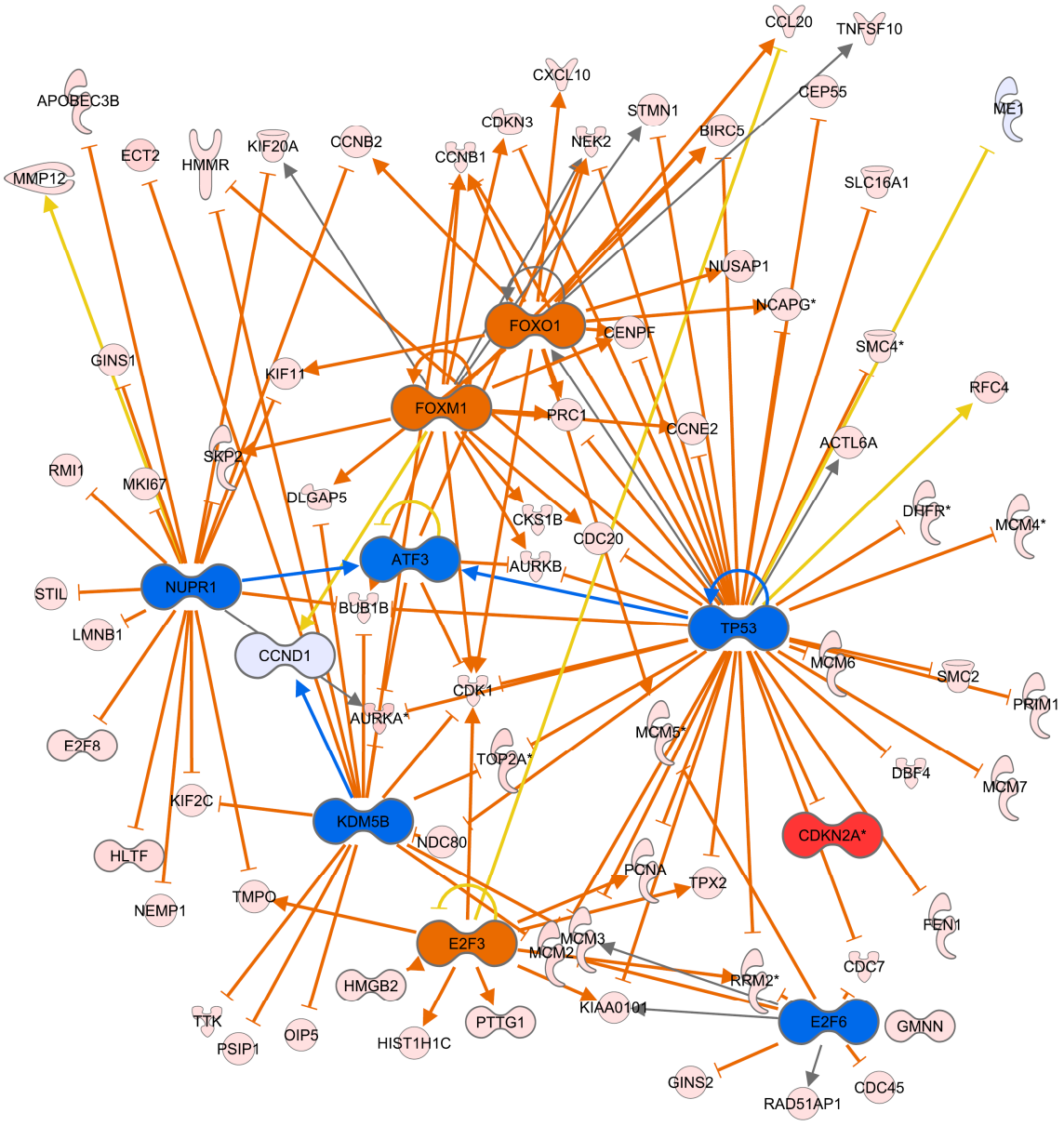


Figure 8: Network of upstream regulators inferred from the metasignature using Ingenuity Pathway Analysis. Blue = inferred inhibition , Orange = inferred activation , Red and Green represent overexpressed and underexpressed genes in HPV+ tumours respectively, ,Solid Line = direct relationships , Dashed Line = indirect relationships, Orange Line = inferred state of upstream regulator is consistent with expression state of effector genes, Yellow Line = inferred state of upstream regulator inconsistent with expression state of effector genes (Appendix A2).

As expected, two of the significantly inhibited upstream regulators are TP53, which is the primary target of E6, and ATF3, which has been implicated as an activator of TP53 by blocking E6-mediated degradation through antagonistic interactions with E6AP (Wang, Mo et al. 2010)**(Figure 8)**.

Interestingly, the signature is suggestive of impaired transcriptional influence of p53 that is caused by its depletion by E6 as opposed to the gain of aberrant transcription in tumours that acquire gain-of-function mutations in *TP53* (Brosh and Rotter 2009). *TP53* mutations are a prominent feature of HPV- HNSCs (Perez-Ordonez, Beauchemin et al. 2006, Stransky, Egloff et al. 2011, Lechner, Frampton et al. 2013), thus the signature may reflect different mechanisms of p53 dysregulation in HPV+ and HPV- tumours. Other dysregulated upstream regulators, as expected from prior discoveries on HPV oncoprotein function, include the inferred activation of E2F3, which is amongst the transcription factors in the E2F family repressed by Rb and the inhibition of E2F6, which is known to be bound by HPV16 E7. (McLaughlin-Drubin, Huh et al. 2008)

In addition to these expected regulators, I also identified several potential regulatory genes not previously implicated in HPV biology, including *KDM5b, FOXO, FOXM1 and NUPR1.* KDM5B, a demethylase responsible for the repression of transcription through removal of trimethyl marks from lysine 4 of histone H3 was identified as being a significantly inhibited upstream regulator. KDM5B has been shown to act in association with pRb in the repression of E2F target genes, thus the inferred downregulation of KDM5B function in HPV+ cancers may be a consequence of pRb targeting by E7(Nijwening, Geutjes et al. 2011).

Other significantly dysregulated upstream regulators in HPV+ tumours include NUPR1, whose function is also inhibited in the HPV+ tumours. This is a multifunctional protein involved in response to cellular stress that has been implicated in chemoresistance to taxols and doxorubicin through the upregulation of p21 and antiapoptotic Bcl-x(L) (Clark, Mitra et al. 2008, Chowdhury, Samant et al. 2009).

FOXM1 was another upstream regulator that was inferred to be activated and this is consistent with its role in modulating the activation of E2F-target genes as part of the DP1-pRb family-E2F DREAM complex in combination with HPVE7 during during G2-M progression (DeCaprio 2014).

Epigenetic modifiers are represented in the metasignature and are putative therapeutic targets.

The epigenetic regulators in the signature include ACTL6a, HTLF and LSH, all chromatin remodellers associated with the SWI/SNF complex, in addition to the histone demethylase EZH2 which has been previously implicated in HPV-driven cancers (Buitrago-Perez, Garaulet et al. 2009). ACTL6a has been implicated as a master regulator of stemness in epidermal keratinocytes (Bao, Tang et al. 2013) and has been shown to be necessary for E6/E7 transcription in CESC cell lines (Lee, Lee et al. 2011) while HELLS has been shown to drive E2F3 (inferred to be activated based on the metasignature) mediated transcriptional programmes (von Eyss, Maaskola et al. 2012); consequently, the metasignature includes multiple novel chromatin remodellers subject to dysregulation in HPV-driven tumours that may be amenable to therapeutic targeting.

Validation of the meta-signature reveals reliable classification performance.

In order to test the performance of the metasignature in an independent dataset, I used RNA-seq profiles for HNSC performed by TCGA. The meta-signature was compared to 10 random sets of 157 features not in the signature and to two other previously published signatures using Kappa as the accuracy metric across 10 iterations of 10 fold Cross-Validation.

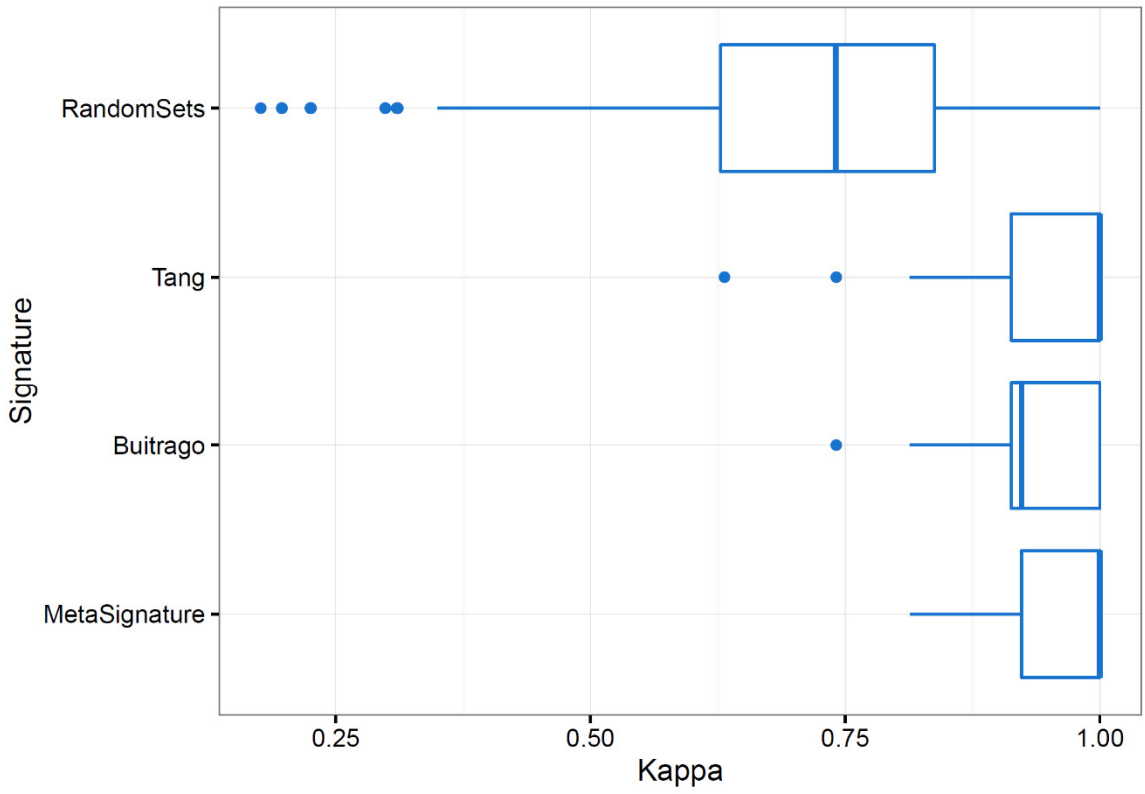


Figure 9: Comparison of Out-of-fold Kappa Values in the TCGA HNSC dataset between the MetaSignature, models trained on 10 random sets of non-metasignature features, and previously published signatures. X-axis = Kappa values, Y-axis = Model.

The signature performed significantly better than the random sets (p-value < 2.2e-16, Wilcoxon's Rank Sum Test) and comparably to a signature derived using list comparison and from supervised analysis from the validation dataset previously reported **(Figure 9)** .

Notably, the metasignature contained new genes not previously implicated as being transcriptionally dysregulated in HPV driven cancers. Visualising the expression of Metasignature genes in the TCGA cohort indicated highly accurate clustering by HPV status **(Figure 10).** HPV+ tumours appeared to be correctly classified even if non-Oropharyngeal, leading to further analyses of subsite specific differences.

A large subset of the Metasignature is uniquely dysregulated in HPV+ tumours.

Having established using the TCGA HNSC cohort that the metasignature is differentially expressed between HPV+ HNSCs, HPV- HNSCs/Normal tissue, I tested if these patterns were unique to HPV+ tumours when considered relative to a broad range of other tumour types arising in a variety of tissues. Using a collection of 13 other tumour types (minimum 5 normal samples) and linear modelling in addition to HNSCs, testing for differential expression of metasignature genes between HPV+ and HPV- tumours after controlling for anatomic site and Cancer/Normal status revealed 80/157 genes to be differentially expressed at 2FC, FDR < 0.01, suggesting many of these transcriptional changes are unique to HPV+ tumours.

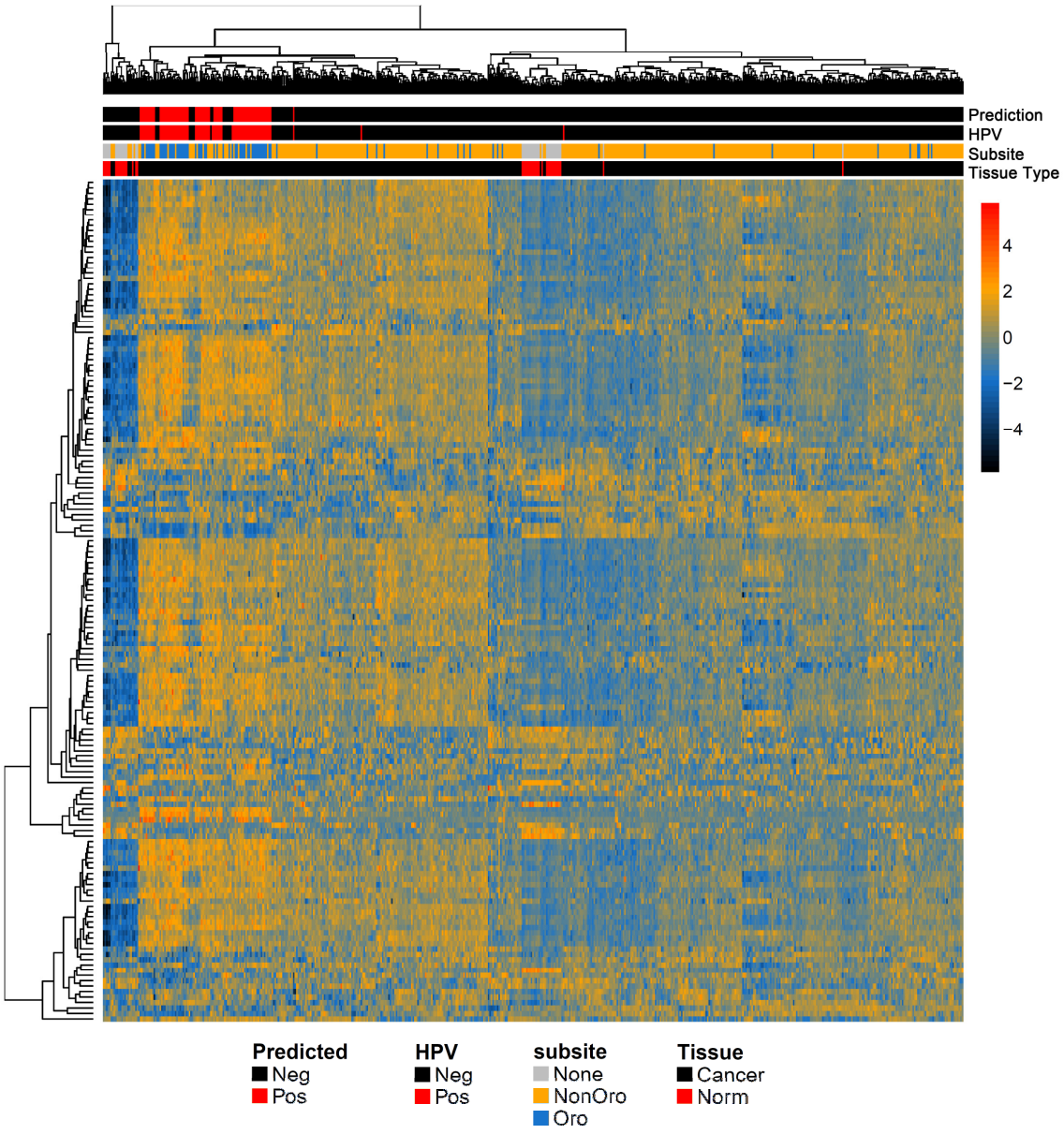


Figure 10: Heatmap of metasignature genes in the TCGA HNSC dataset (n = 561). Samples in columns and expression Z-scores in rows. Annotation rows from top to bottom indicate Random Forest predictions, HPV status by RNA-transcripts, anatomical subsite and Cancer/Normal status. Neg = HPV- , Pos= HPV+ , Oro = Oropharyngeal, NonOro = Non-Oropharyngeal, Norm = Normal.

HPV transcript-positive OPSCC and non-OPSCC share common transcriptomic/methylomic patterns.

The hypothesis that HPV+ non-OPSCCs share HPV-associated molecular profiles with HPV+ OPSCCs was tested on RNA-sequenced HNSC from the TCGA (n=520, 54 HPV+ OPSCC, 21 HPV+ Non-OPSCC, 26 HPV- OPSCC, 419 HPV- non-OPSCC) using the Random Forest classifier trained on the expression signature as described earlier in this chapter and also one trained on a signature of Methylation Variable Positions (MVPs) different between HPV+ and HPV- OPSCC.

Heatmapping revealed that in both cases **(Figure 11)**, HPV+ tumours tended to cluster together into one or two distinct clusters, with OPSCC and non-OPSCC samples interspersed. Analyses of classification using aggregated out-of-fold predictions from Random Forests showed robust classification performance and the appropriate classification of the vast majority of samples, with 71/75 HPV+ tumours correctly classified by the Expression-based classifier (Kappa = 0.96) and 61/68 HPV+ tumours correctly classified by the Methylation classifier (Kappa = 0.92).

Having demonstrated that transcriptional and epigenomic patterns are conserved between HPV+ OPSCCs and Non-OPSCCs, expression patterns of the primary viral oncogenes E6 and E7 were compared by anatomical subsite. Overall distributions for both E6 and E7 were similar and no significant differences in distributions were identified (Wilcoxon's Rank Sum Test) **(Figure 12)**.

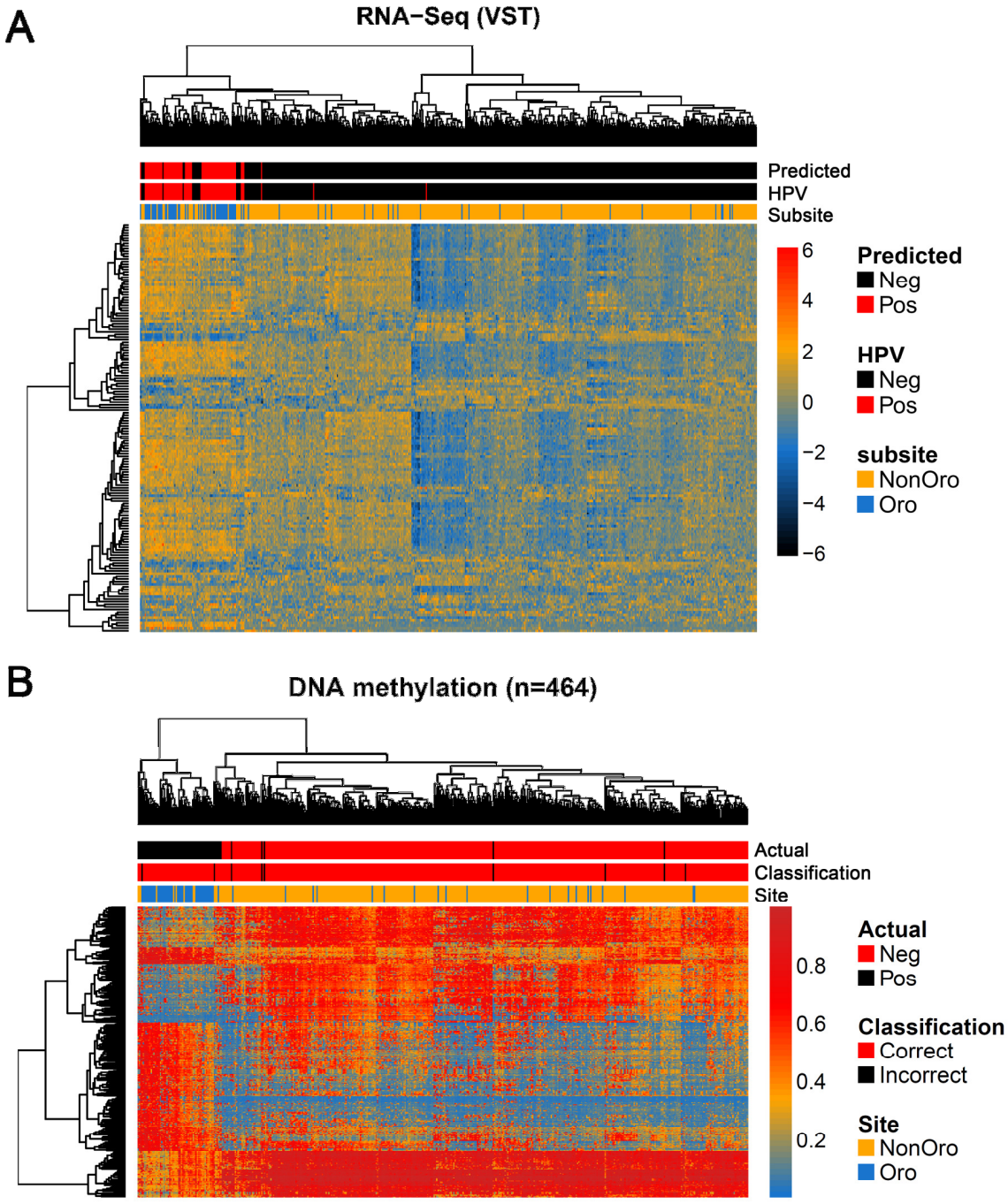


Figure 11: Visualisation of HPV-associated molecular signatures in the TCGA HNSC cohort. A) Expression metasignature (n=520), B) MVPs associated with HPV+ OPSCC (Oro) relative to HPV- OPSCC are also preserved in HPV+ non-OPSCC (non-Oro) . Annotation bars show Random Forest predictions, HPV transcript status and anatomic subsite. Samples in columns and features in rows, with Z-scores represented for expression data and beta-values for methylation.11A is reproduced from Figure 10 for convenience.

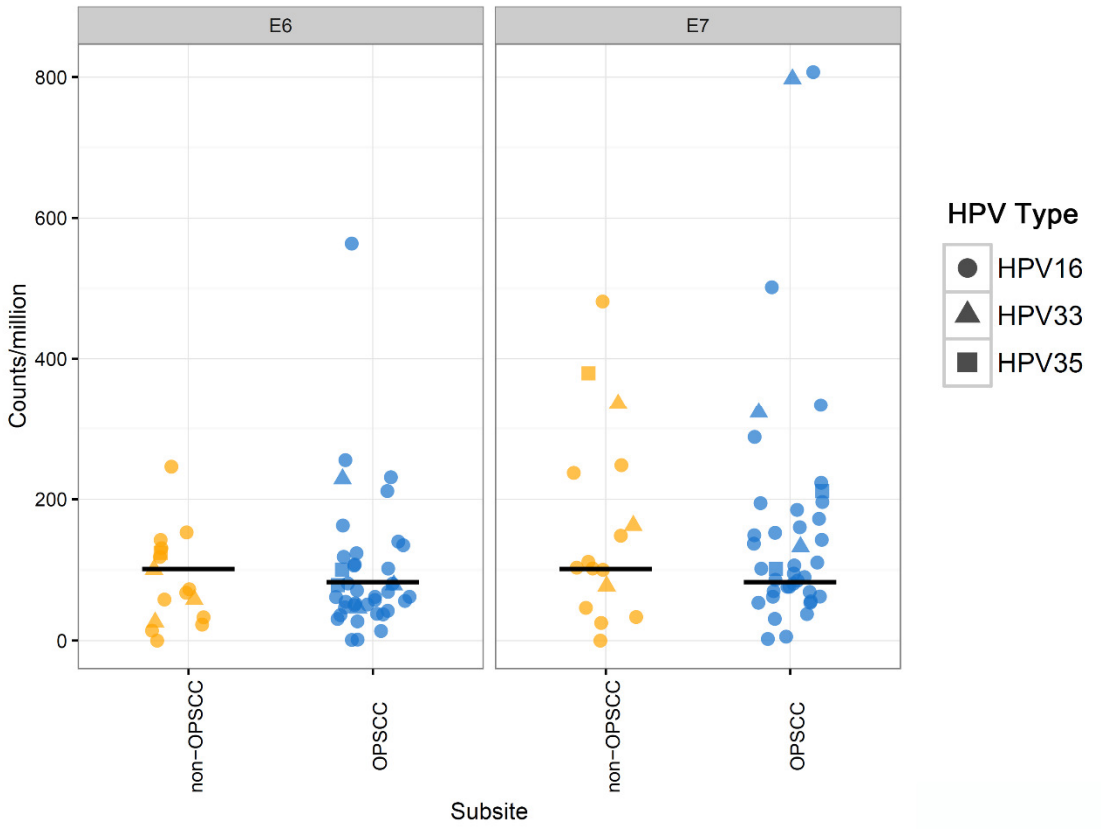


Figure 12: Distributions of HPV E6 and E7 expression are similar between HPV+ OPSCC and non-OPSCC for HPV types found in both subsites.

Genomic analyses further support an aetiological role for HPV outside the Oropharynx.

Having observed strong similarities between HPV+ tumours from oropharyngeal and non-oropharyngeal sites at the gene expression and epigenetic levels, I investigated if large-scale genomic similarities existed by subsite.

Previous work has revealed that HPV- and HPV+ OPSCCs display characteristic mutations and/or copy number alterations (CNAs). In particular, alterations to *TP53, CDKN2A* and *CCND1* are almost exclusive to HPV-negative OPSCC, in which they occur at high frequency ((Gillison, Chaturvedi et al. 2015), (van Houten, Snijders et al. 2001), (Lechner, Frampton et al. 2013), (TCGA 2015)). In HPV-driven OPSCC and cervical cancer, it is thought that E6-mediated p53 degradation removes the selection pressure for TP53 gene mutations and that bypass of the G1/S checkpoint by E7 likewise obviates the pressure to acquire CDKN2A or CCND1 alterations (Westra, Taube et al. 2008, Lechner, Frampton et al. 2013).

Group	TP53 mutations	CDKN2A mutations	CCND1 Amplification	CDKN2A Deletions (Deep + Shallow Deletion)
HPV + OPSCC	0/50 (0%)	0/50 (0%)	6/54 (11.1%)	3 + 1 /54 (7.4%)
HPV+ Non-OPSCC	1/18 (5.5%)	0/18 (0%)	3/21 (14.2%)	1 + 2/21 (14.2%)
HPV- OPSCC	23/25 (92%)	5/25 (20%)	19/26 (73.7%)	10+12/26 (84%)
HPV- Non-OPSCC	336/409 (82%)	107/409 (26%)	197/413 (47.6%)	140+141/413 (68%)

Table 6: Breakdown of Hallmark genomic alterations in CDKN2A, TP53 and CCND1. HPV+ OPSCC and Non-OPSCC display similar frequencies of alteration.

As expected, high frequencies of *TP53* mutations, *CDKN2A* mutations, *CDKN2A* deletions and CCND1 amplifications were evident in both HPV- OPSCC and non-OPSCC but not in HPV+ OPSCC. The HPV+ non-OPSCC tumours resembled the HPV+ OPSCCs (**Table 6**). Taken together, these genomic analyses further buttress the characterisation of HPV as a driver in transcript-positive non-OPSCC.

Of the 3 HPV+ tumours that were misclassified as HPV- on the basis of both the transcriptional and epigenetic signatures, two displayed low viral transcript counts of 125 and 5908 viral reads (overall mean=35204; range 125-202666), one had a *CCND1* amplification and a *CDKN2A* deletion (much more common in HPV- disease) and one contained an atypical HPV-type that was a singleton in the dataset (HPV56). This suggests that even the presence of HPV RNA may occasionally lead to erroneous conclusions about aetiology - highlighting the importance of an integrated molecular approach to identifying the HPV status of these tumours. The 3 misclassified samples were removed for further analysis.

HPV+ HNSC may show subsite-associated prognostic differences.

Large clinical trials have established that HPV+ OPSCC patients have markedly better clinical outcomes compared to HPV- OPSCC (Fakhry, Westra et al. 2008, Ang, Harris et al. 2010) . These findings have led to de-escalation trials aimed at reducing the treatment associated morbidity seen with chemoradiation when Oropharyngeal cancers are HPV-driven (reviewed in (Vokes, Agrawal et al. 2015)). The discovery that molecular profiles are conserved between HPV+ OPSCC and non-OPSCC raises questions of clinical behaviour and management in the latter.

To date, two previous studies have investigated the relationship between p16 overexpression, (a proven biomarker for HPV status in OPSCC) and outcome in non-OPSCCs. One reported no survival benefit associated with p16 expression outside the oropharynx (Lassen, Primdahl et al. 2014) but did so in OPSCC, whereas analyses from another trial showed that p16-overexpressing tumours were associated with better outcomes regardless of anatomical subsite, but did not demonstrate benefits associated with HPV DNA positivity in non-OPSCCs (Chung, Zhang et al. 2014).

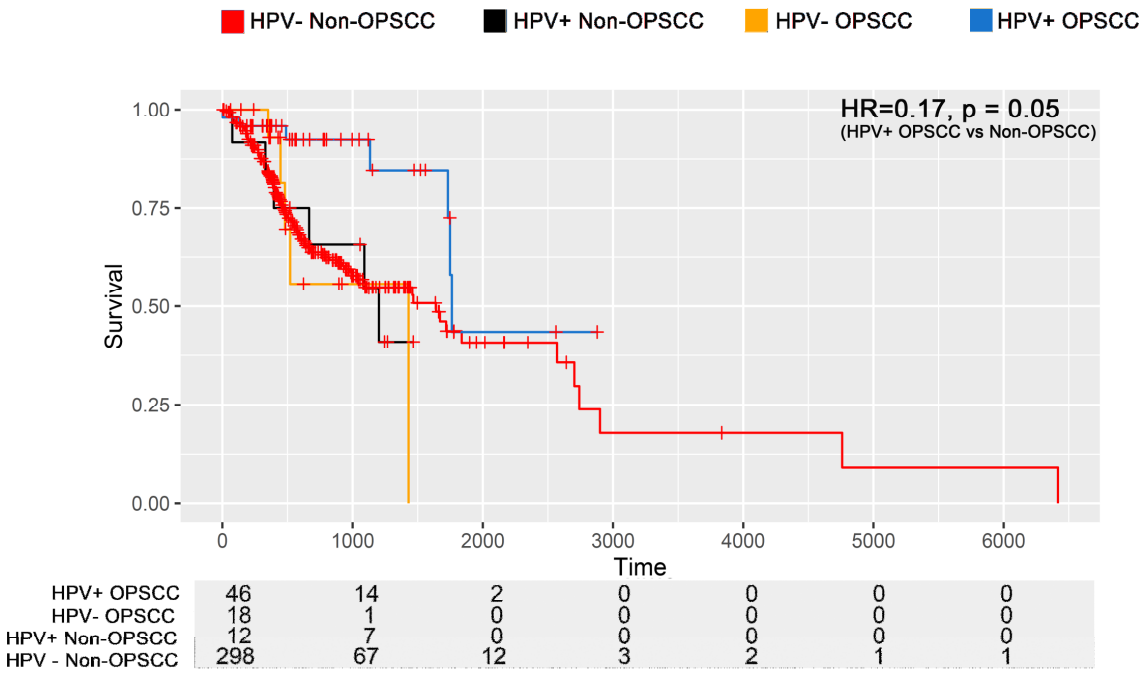


Figure 13: Overall Survival in the TCGA HNSC cohort, stratified by HPV status and anatomic subsite. HPV+ OPSCC was associated with significantly better survival than HPV+ non-OPSCC. P-value and Hazard Ratio from multivariate Cox regression. X-axis = time in days, Y-axis = survival probability. The table indicates number of patients at risk of death.

Kaplan-Meier curves **(Figure 13)** showed that HPV+ OPSCC and HPV+ non-OPSCC may be associated with different clinical trajectories (HR = 0.17, p = 0.05, CI = 0.03 – 0.99). This led me to consider various factors to explain this prognostic gap by subsite.

Firstly, comparisons were carried out of the transcriptomes of HPV+ OPSCC vs HPV+ Non-OPSCC, leading to the identification of 68 DEGs (FDR < 0.01, 2FC). Pathway analyses of this gene set identified "migration of cells" as the most enriched process in the HPV+ non-OPSCC tumours (Activation z-score = 2.69, p = 2.46E-03) attesting to an additional layer of transcriptional dysregulation as one putative reason for survival differences by anatomic subsite in HPV+ HNSC. Comparisons of methylation data showed no MVPs significantly different (delta-Beta > 0.4, FDR < 0.01). Moreover, almost universal retention of wild-type TP53 (previously linked to increased radiosensitivity in HPV+ OPSCC (Kimple, Smith et al. 2013)) in HPV+ non-OPSCC **(Table 1)** was observed, suggesting an alternative explanation for the survival difference by subsite within HPV+ tumours.

Anatomic subsite is associated with differences in TIL levels and activity

Reviewing the literature surrounding determinants of outcome in HPV+ tumours, revealed that levels of tumour-infiltrating lymphocytes (TILs) are correlated with outcome in HPV+ OPSCC, with low TIL numbers designating a poor-prognosis group of HPV+ patients whose survival did not differ from patients with HPV- OPSCCs (Ward, Thirdborough et al. 2014). Immune infiltration has also been described as a prognostic stratifier in HPV-associated anal cancer (Gilbert, Serup-Hansen et al. 2016).

The importance of the immune system in shaping the evolution of HPV+ OPSCC is also reflected by HPV+ OPSCC cells frequently expressing high levels of the immune checkpoint ligand PD-L1 to evade immune destruction (Lyford-Pike, Peng et al. 2013). The fact that the tonsils and base of tongue are lymphoid tissues, and that high TIL levels are a common feature of these tumours led to the hypothesis that TIL levels in OPSCCs are higher than those found in non-OPSCCs, allowing a potent therapy-primed immune response in the Oropharyngeal setting and potentially explaining the difference in outcome.

This hypothesis was tested using two measures of infiltrating lymphocytes: firstly, the measurement of CD4 (Helper / Regulatory T lymphocyte), CD8A/CD8B (Cytotoxic T lymphocyte) RNA transcript abundance (VST) as surrogates for immune infiltrate activity led to the discovery of significantly higher levels of these infiltrating lymphocyte markers in HPV+ OPSCCs (All at FDR < 0.05) **(Figure 14A)**.

This analysis was complemented by pathological evaluation of images from H&E stained sections (available from the TCGA Digital Image Archive) by a pathologist (Prof Gareth Thomas, University of Southampton, a collaborator), who was blinded to anatomic subsite. Statistical analysis of his histopathological estimates suggested increased TILs in HPV+ OPSCCs versus HPV+ non-OPSCCs (**Figure 14B**, Cochran-Armitage Trend Test, $p = 0.046$).

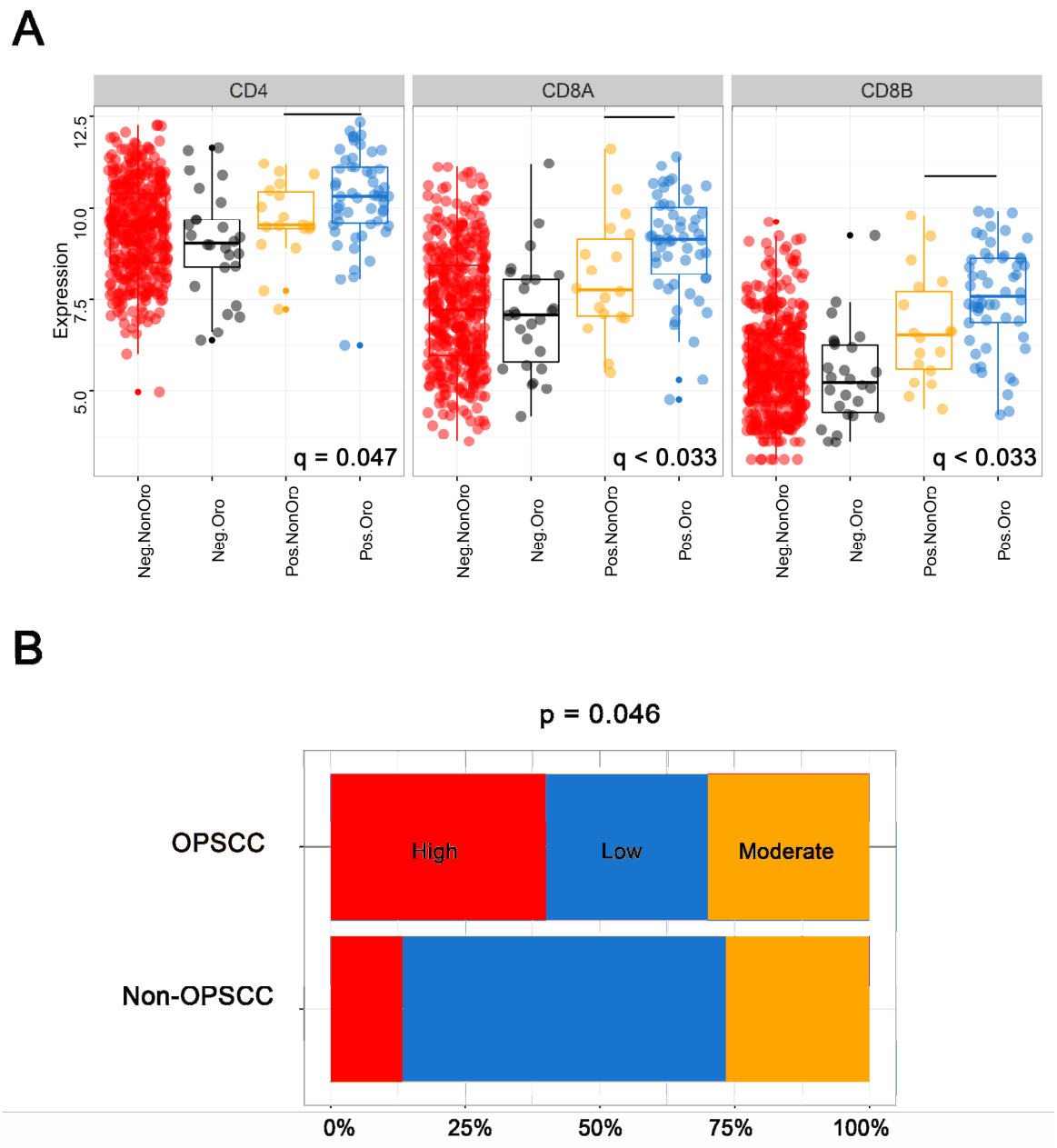


Figure 14: A) Breakdown of TIL receptor expression (log2 counts per million) by combinations of HPV status and anatomical subsite. FDR from Wilcoxon's Rank Sum Test, HPV+ OPSCC vs Non-OPSCC. B) HPV+ OPSCC are more likely to be TIL-high relative to HPV+ Non-OPSCC as scored by a pathologist based on examination of H&E slide images from TCGA. P.Value from Cochran-Armitage test for trend. Pos and Neg represent HPV status and Oro and Non-Oro represent OPSCC and Non-OPSCC respectively in the figure.

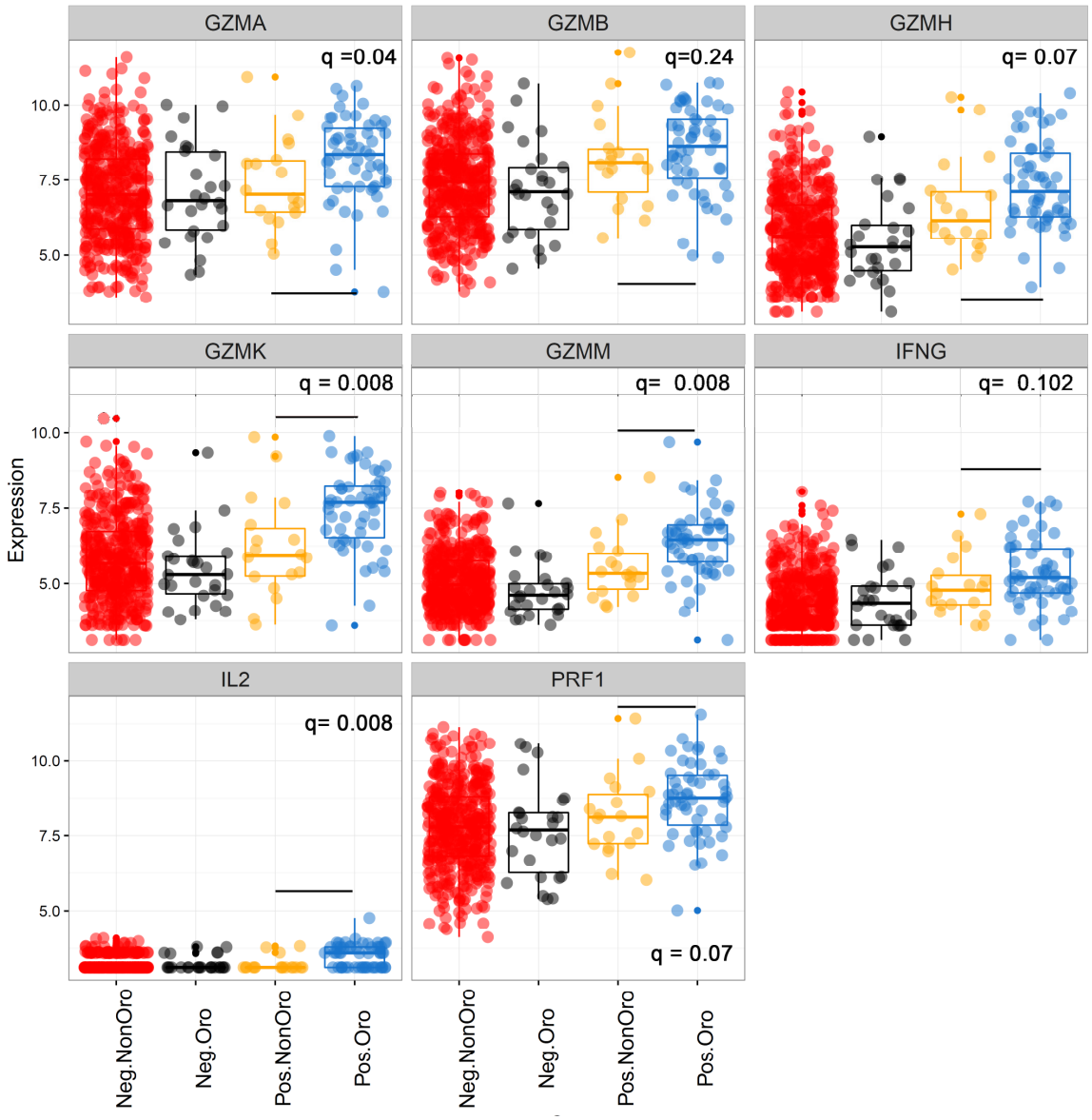


Figure 15: mRNA expression (log2 counts per million) of TIL Effector Molecules. q-values from Wilcoxon's Rank Sum Tests between HPV+ OPSCC and HPV+ non-OPSCC. X axis represents groups by HPV status and anatomic subsite. Pos and Neg represent HPV status and Oro and Non-Oro represent OPSCC and Non-OPSCC respectively in the figure.

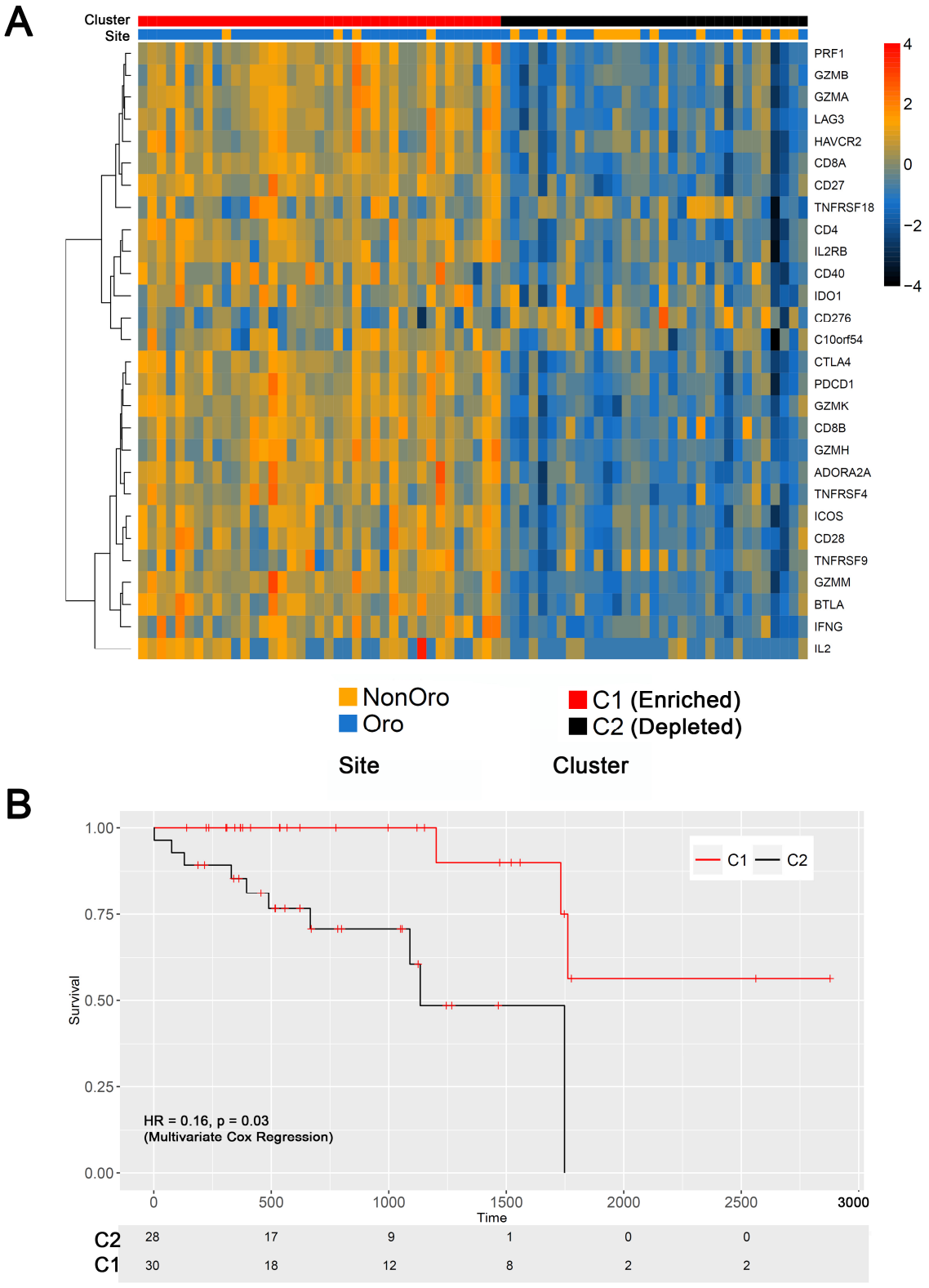


Figure 16: A) Clustering using a panel of immune cell markers, effector genes and immune checkpoint transcripts separates HPV+ HNSCs (Non-Oro = Non-OPSCC, Oro = OPSCC) into two distinct groups. B) These groups show differential survival. Stats from multivariate Cox regression. Y-axis – probability of survival, X-axis – time to most recent followup in days. Table indicates numbers of patients at risk. Also see Appendix A4 for all coefficients.

It was then tested if differences in TIL levels within these subgroups of tumours defined by HPV status and anatomic subsite was associated with differential TIL activity by evaluating the expression of the effector markers granzymes (Granzyme A,B,H,K and M), perforin, their upstream activator IL2 (Janas, Groves et al. 2005), and Interferon-gamma (Bold and Ernst 2012) expression.

HPV+ OPSCC displayed elevated levels of a large subset of these markers relative to each of the other three groups **(Figure 15)** (FDR < 0.1 between HPV+ OPSCC and non-OPSCC). Finally, PAM consensus clustering was carried out using a panel of these markers, effectors and known immune-checkpoint molecules, derived by manual curation. This separated samples into two distinct clusters, marked by high and low expression of the transcripts in the panel respectively **(Figure 16A)**. The low expression cluster showed enrichment for HPV+ non-OPSCC (14/18) relative to HPV+ OPSCC (19/54) (p = 0.002, Fisher's exact test) further reflective of low immune infiltrates. Immune cluster was then found to be significantly associated with outcome (immune cluster high vs low HR=0.1622, p=0.0431, CI=0.035-0.8595, **Figure 16B**) after controlling for anatomic subsite, T-stage, node stage and age at diagnosis (NB – these covariates have all been shown to be relevant within HPV+ OPSCC, see methods for dichotomisation criteria). Secondary validation based on clustering samples using four WGCNA (Weighted Gene Correlation Network Analysis) modules previously associated with HPV+ tumours, some of which were also correlated with lymphocyte activity (Ottensmeier, Perry et al. 2016) also recapitulated similar associations with survival in multivariate analyses (HR = 0.21, p = 0.031, CI=0.05-0.86). These findings demonstrate that HPV+ OPSCCs are subject to

higher levels of TIL activity, and display an immunoactive transcriptional profile that is prognostically informative.

Intriguingly, both of the gene sets used to define the clusters above show inexact correspondence with histopathological estimates of lymphocyte infiltration; using the immune activity panel I derived in this chapter, 7/18 High-TIL, 5/10 Moderate-TIL and 16/20 low-TIL samples were associated with the low-expression cluster. This indicates there may be samples with high levels of infiltration that nonetheless are not marked by effector function, or that there may be tumours with small numbers of infiltrating lymphocytes that may still be capable of potent anti-tumour responses, marked by distinct molecular profiles.

Chapter Conclusions

The work carried out in this chapter represents, to date, the most comprehensive assembly and analysis of transcriptomes from models and tumours transformed by HPV. Transcriptional patterns that mark HPV+ tumours are consistent with origins in the direct activity of HPV E6 and E7, which is evident in the induction of the signature in transformed Mesenchymal Stem Cells, with marked changes in expression patterns of genes that regulate cell cycle progression. Validation of the metasignature through machine learning and comparisons using RNA-seq data across a large panel of cancer types both indicate that the metasignature is capable of being a functional readout for the activity of HPV.

A large subset of the metasignature is perturbed in unique ways and to magnitudes not typical of many other tumour types and is of potential utility for developing transcriptional biomarkers to test for the active involvement of HPV in tumourigenesis.

The confirmation of a driver role for HPV in non-OPSCC has multiple implications for clinical management of these tumours, firstly, the worse prognosis associated with these tumours suggests therapeutic de-escalation is ill-advised, and secondly, the generally immune-depleted pattern associated with these tumours suggests that HPV+ OPSCC patients may be more likely to respond to immune-checkpoint blockade relative to HPV+ non-OPSCC, while drugs that specifically target HPV-driven tumours may be useful regardless of subsite. The immune activity transcriptional profiles derived in this chapter will facilitate the stratification of patients for immunotherapy.

Chapter 4: APOBEC-mediated mutagenesis is a key driver of genomic evolution in HPV+ cancers.

Note – interim analyses of the work presented in this chapter were published as a paper (Henderson, Chakravarthy et al. 2014) involving collaborative software development and analyses with Stephen Henderson of the Bill Lyons Informatics Centre, UCL Cancer Institute. I received joint-lead authorship for my contribution to the interim analysis. I generated all new code where required and performed all the analyses on the extended dataset used in this chapter, while leveraging on the tools and ideas established and generated during the interim analysis.

Multiple genes in the gene expression signature suggest involvement of APOBEC-activity in mutagenesis and genomic evolution.

One of the many metasignature genes that were also unique to HPV+ tumours across the pan-cancer cohort was *APOBEC3B* (Fold change 3.48), which encodes a cytidine deaminase enzyme responsible for editing of viral DNA during the innate immune response to infection. Mutations indicative of APOBEC editing have been reported in both HPV DNA isolated from cervical dysplasias and more recently in cellular DNA from HPV+ HNSC and CESC; presumably the result of an off-target activity that can be selected for during tumour development (Vartanian, Guetard et al. 2008, Burns, Temiz et al. 2013, Roberts, Lawrence et al. 2013, Henderson, Chakravarthy et al. 2014).

When expression was visualised across samples from the TCGA PanCan dataset for cancer types for which at least 5 normal samples were available, HPV+ HNSCs were found to have amongst the highest expression of *APOBEC3B* **(Figure 17)**. APOBEC3B targets ssDNA, and it has been proposed that exposure of the ssDNA substrate, as occurs at stalled replication forks, rather than the level of APOBEC enzymes, is the key limiting factor for this mode of somatic mutagenesis (Roberts and Gordenin 2014).

Findings presented in the previous chapter indicate that HPV+ tumour cells express multiple genes involved in the repair of DNA damaged during replication and together with the known ability of E6 and E7 to induce replication fork stalling, suggested that HPV-driven tumourigenesis might be associated with APOBEC-mediated mutagenesis.

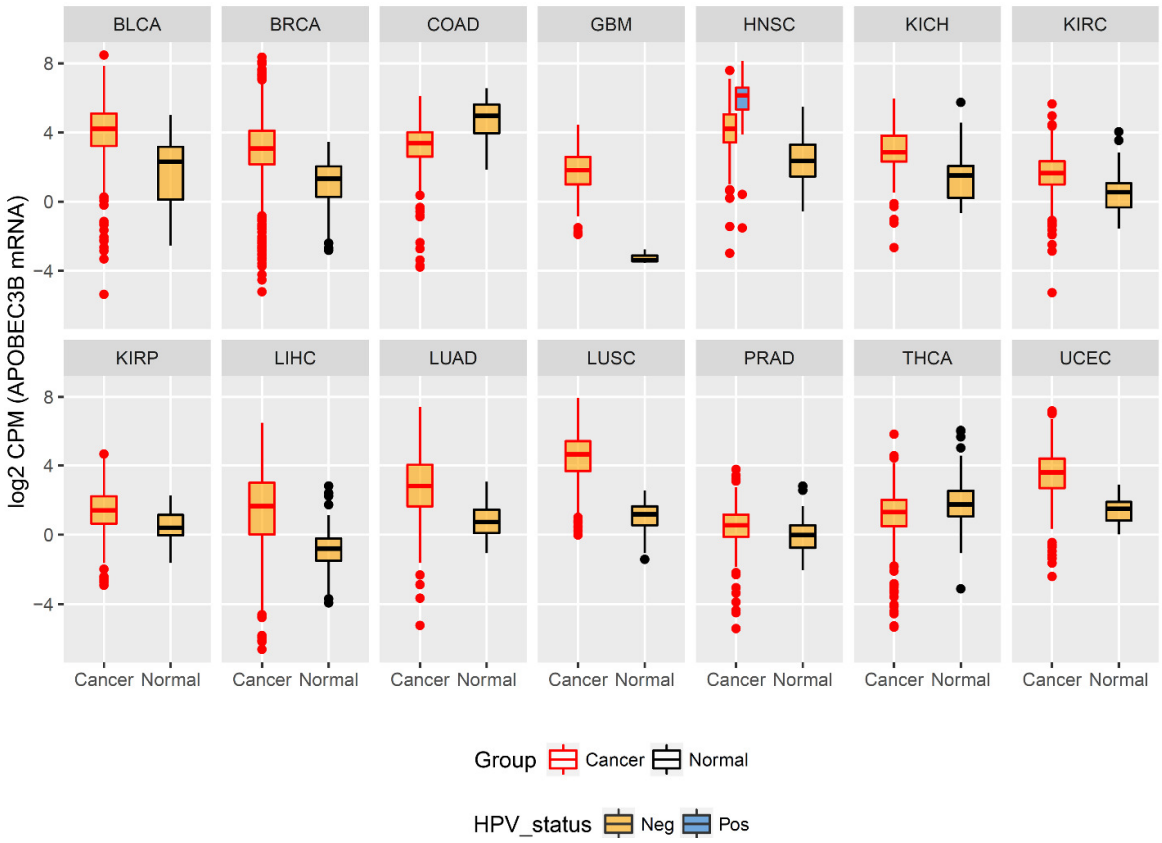


Figure 17:Plot of *APOBEC3B* expression (log2 (cpm+0.5)) across tumour types with both normal and cancer samples available from the TCGA (minimum 5 normal samples). HPV+ HNSCs (red) exhibit some of the highest levels of APOBEC3B expression.

APOBEC mediated mutagenesis is significantly enriched in HPV+ HNSC relative to HPV- HNSC.

While previous reports indicated a very high level of APOBEC-induced mutations in cervical tumours, it was unclear if this was associated with HPV or with tissue of origin. HPV-association was therefore tested by comparing HPV+ and HPV- HNSCs.

Previously, four distinct mutational signatures that were defined across a broad cohort of cancer genomes and exomes were found to be operative in HNSC, with APOBEC implicated in generating two of these. Using a method to deconvolute spectra (Rosenthal, McGranahan et al. 2016) into signatures on a per-sample basis, the number of mutations mapping to these signatures and those that did not were estimated.

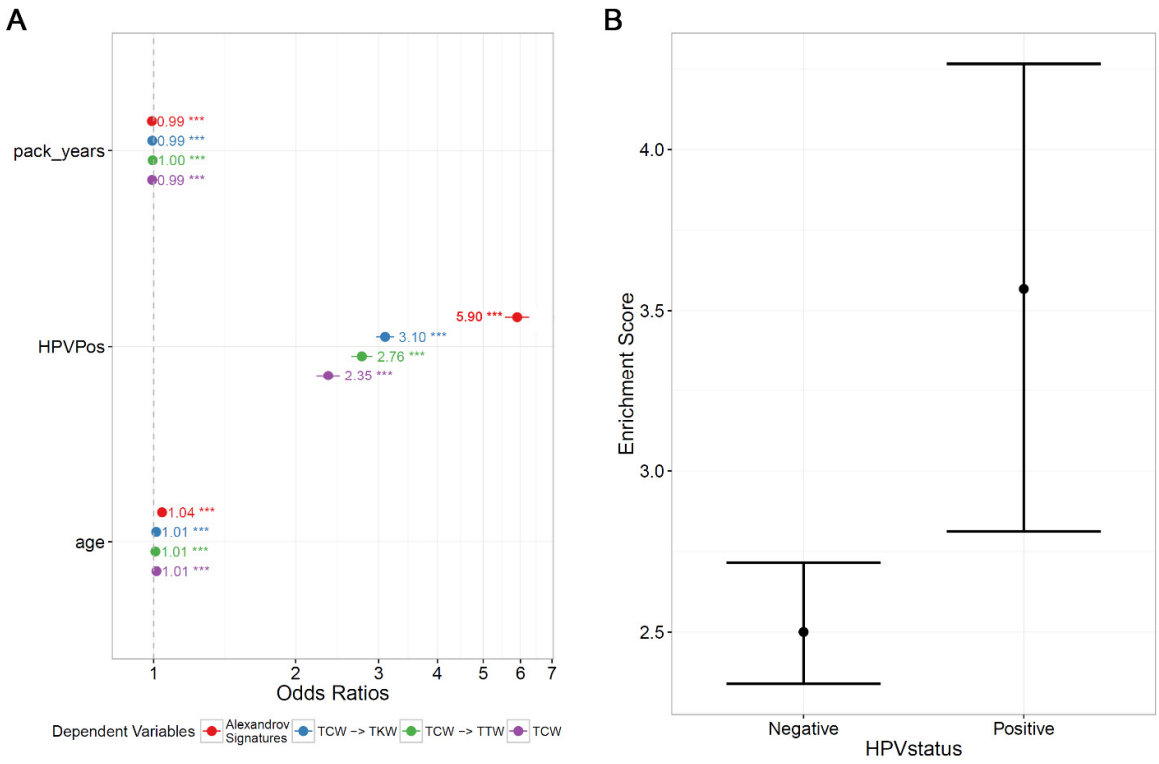


Figure 18: A - GLM analysis reveals significant associations with HPV status for APOBEC signatures according to the proportion of mutations categorised as Alexandrov APOBEC mutations or supervised analysis using subclasses of TCW -> TKW mutations. B - calculating localised enrichment scores for APOBEC to non-APOBEC mutations reveals significant differences in the distributions of APOBEC-enrichment scores by HPV status in HNSCs.

Fitting binomial generalised linear models to the proportion of mutations attributable to the Alexandrov APOBEC unsupervised signature (comprised of signatures 2 and 13 from (Alexandrov, Nik-Zainal et al. 2013), representing both C>T mutations that are associated with multiple processes and C>G mutations that are more APOBEC-specific) as well as

proportions of APOBEC mutations defined by supervised analysis using TCW to TKW mutation fraction as a determinant of APOBEC mediated mutagenesis, where K = T/G and W=A/T identified HPV status was a significant predictor of candidate APOBEC-mediated mutations with a large effect size (OR range across models – 2.35 – 5.90) **(Figure 18A)** while age and smoking resulted in small and significant effects, which is not surprising since they are processes that contribute to mutational signatures that are non-APOBEC.

These approaches were also validated by comparison to a localised enrichment score based on the likelihood of mutations based on the local sequence context of mutant bases and HPV+ tumours were again significantly enriched upon permutation testing of the difference in means (p=0.0045) (**Figure 18B**). In all cases, HPV status was a strong predictor of enrichment for APOBEC-mediated mutagenesis.

APOBEC-mediated mutagenesis is not part of a generalised antiviral response.

Given the role of APOBEC-mediated cytosine deamination in antiviral immunity, I then tested if other virally-driven tumours were similar to HPV+ tumours (194 CESC, 68 HPV+ HNSC) in displaying enrichment for this mutational process using HBV/HCV+ (Hepatitis B Virus/ Hepatitis C Virus) Hepatocellular carcinomas (HCC, n=213) and EBV+ Stomach

Adenocarcinomas (STAD, n=26).

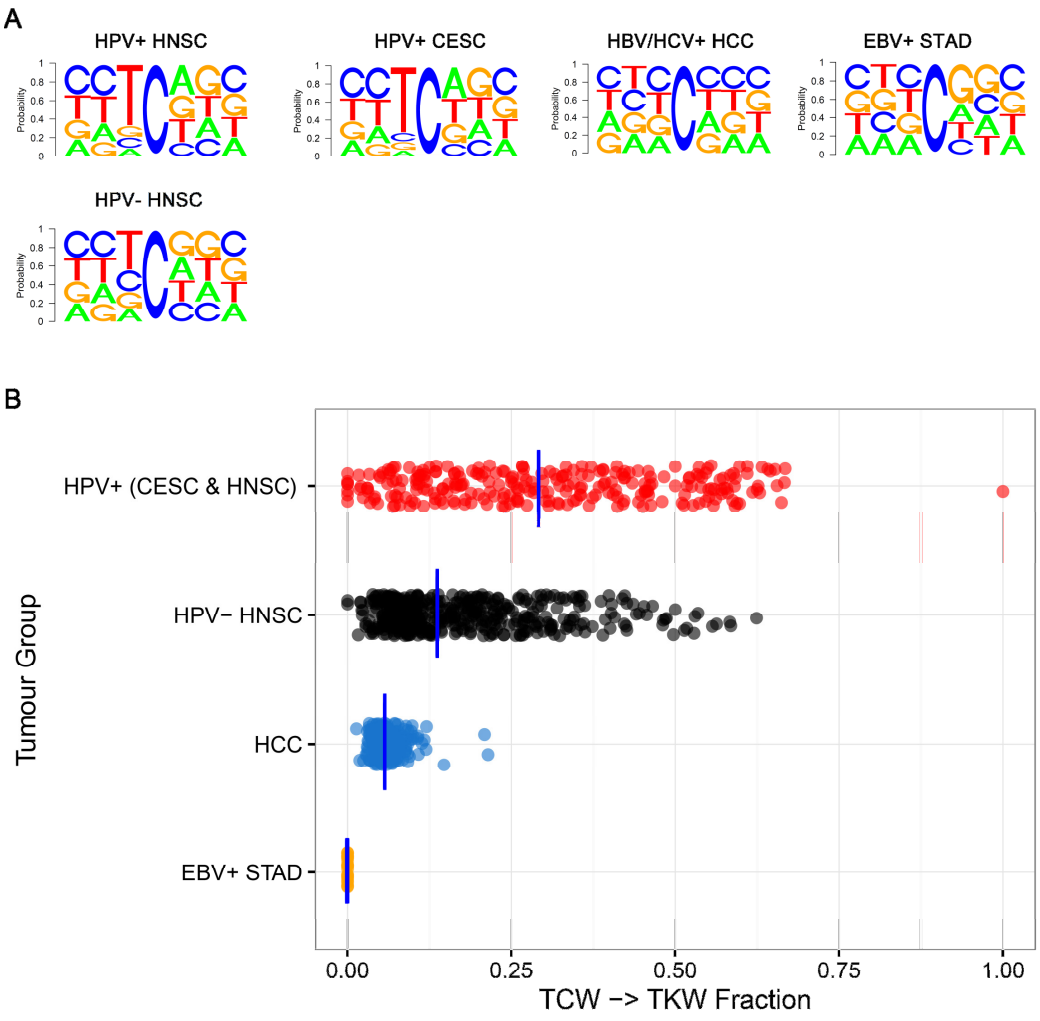


Figure 19: A) Sequence logos for tumour subsets. HPV+ HNSC, HPV+ CESC, and even HPV- HNSC display a TCW motif in their sequence logos relative to Hepatocellular carcinomas or EBV+ Stomach Adenocarcinomas. B) Breakdown of TCW -> TKW fractions by tumour subset. HPV+ tumours are greatly enriched relative to viral HCC or EBV+ Stomach Adenocarcinomas, suggesting APOBEC mediated mutagenesis is not a general feature of virally driven cancers. Crossbars represent group medians.

Visualisation of sequence logos showed markedly lower levels of enrichment in non-HPV viral tumours **(Figure 19A)**, and comparing distributions of TCW -> TKW fractions **(Figure 19B)** revealed significantly higher medians in HPV+ tumours (0.29) relative to HCC (0.05, q.value < 2e-16, Wilcoxon's Rank Sum Test) and STAD (0, q.value < 3.4e-16, Wilcoxon's Rank Sum Test). This indicates APOBEC-mediated mutagenesis is not a generalised antiviral response that marks the genomic evolution of virally driven cancers.

Enrichment for APOBEC-mediated mutagenesis is retained across mutational profiles of candidate driver mutations

The genomes of cancers are punctuated by a wide range of mutations, some which have an impact on the fitness of cancer cells (drivers) and those that play a passive role (passengers). Driver mutations show distinct clustering and recurrence patterns allowing statistical classification, as implemented in the MutSig CV suite of tools.

The exome-wide enrichment of APOBEC-mediated mutagenesis seen in HPV+ tumours raised questions of functional significance; whether TCW -> TKW mutations merely served to generate passenger mutations or played a critical role even in generation of driver mutations. This was tested by comparing distributions of TCW->TKW mutations in MutSig CV associated genes and distributions of TCW -> TKW mutations from exome-wide sampling that was blind to being accorded driver-status by MutSig. Visual observation **(Figure 20)** and statistical testing using Wilcoxon's Rank Sum Test suggested that distributions and differences between HPV+ and HPV- tumours were preserved within the MutSigCV gene subsets.

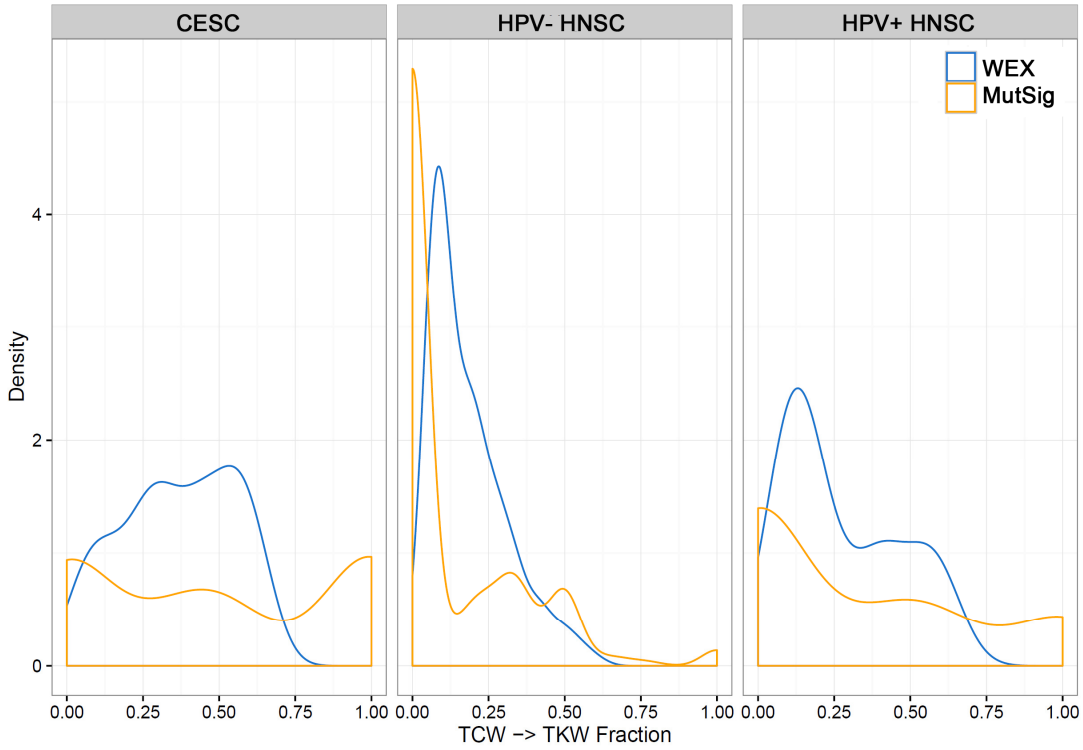


Figure 20: Distributions of TCW -> TKW fractions within whole exomes and MutSig genes by category. X axis = TCW -> TKW fraction. Y axis = density

APOBEC-mediated mutagenesis can determine mutational profiles and hotspot preference in cancer drivers.

The analysis of genes most targeted by APOBEC mediated mutagenesis across the cohort of HPV+ tumours (HNSC and CESC) and HPV- HNSCs serving as a control highlighted *PIK3CA* as one of the most commonly mutated by APOBEC (2nd most common in CESC, 3rd most common in HPV+ HNSC, and 3rd most common in HPV- HNSC). In HPV+ tumours, where APOBEC activity is enriched, it was found that out of 78 *PIK3CA* mutations, 60 were TCW -> TKW, whereas in HPV- tumours, only 35/74 were of this type. *PIK3CA* encodes the p110alpha catalytic subunit (a lipid kinase) of a Phosphoinositide-3-kinase that is activated by Receptor Tyrosine Kinases and RAS. Along with a kinase domain, this protein also contains a helical domain that activates this protein by interacting with a p85 subunit (Liu and Roberts 2006).

Visualising the locations of these mutations suggested they mapped to two distinct mutational hotspots, skewed strongly towards APOBEC mediated mutagenesis in the HPV+ tumours relative to HPV- HNSCs ($p = 0.0003$, Chi-squared test for equality of proportions) **(Figure 21A)**. Mutations in the helical domain (Exon 9) showed enrichment for TCW -> TKW mutations whereas those in the kinase domain (Exon 20) did not. Examining the hotspot mutations, the helical domain hotspot mutations (c.1624G>A and c.1633G>A) were found to be marked by TCW sites on the opposite strand, marking them out as driver mutations potentially driven by APOBEC mediated mutagenesis, whereas the kinase domain mutation hotspot (H1047R) was not **(Figure 21B)**.

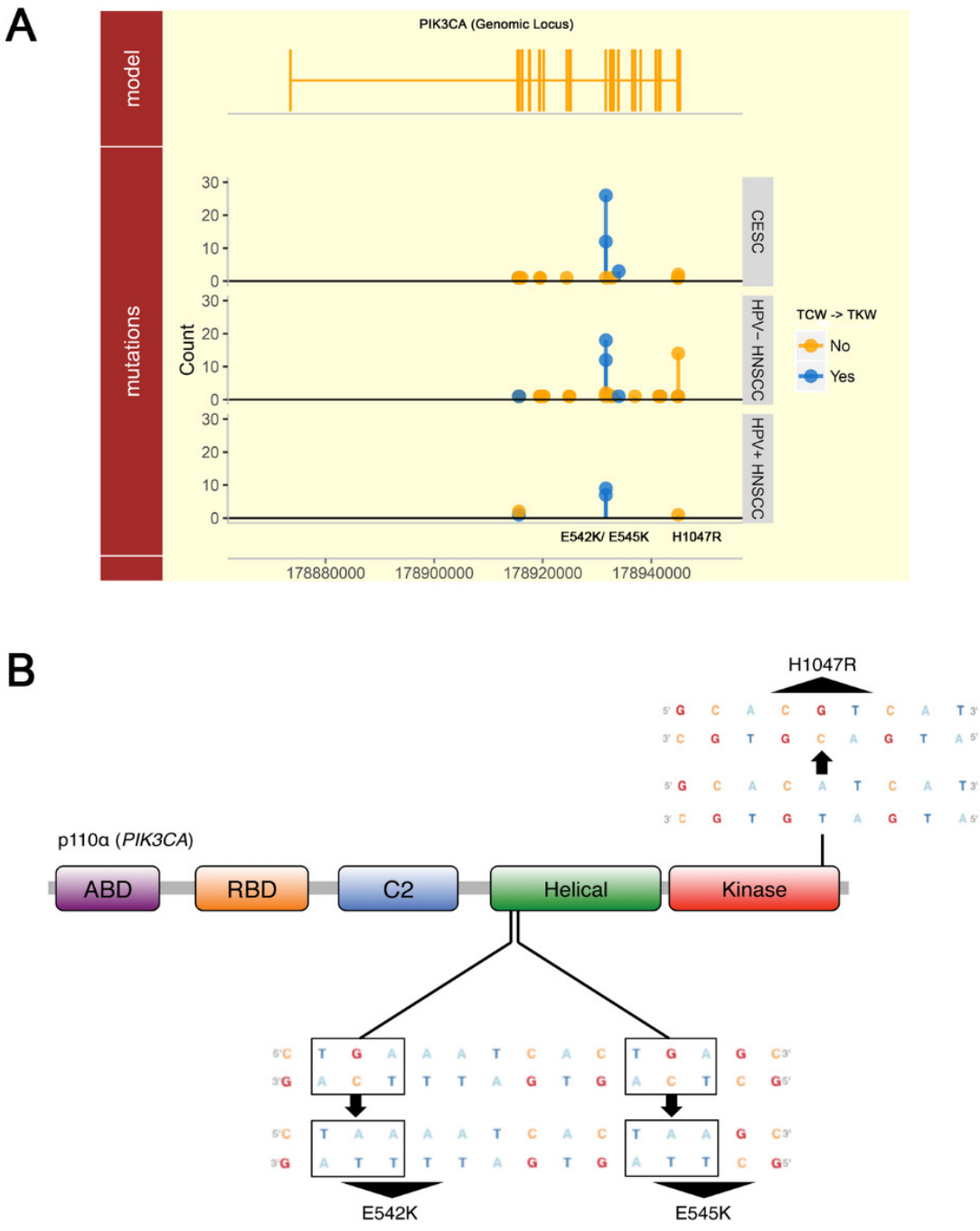


Figure 21: Illustration of PICK3CA hotspots and sequences at hotspots. Sites of helical hotspot mutations harbour a recognition site for APOBEC3B (TCA) on the non-coding strand. Figure 21B from (Henderson, Chakravarthy et al. 2014).

PIK3CA helical hotspot and kinase hotspot mutations tend to be equally common when examined across cancers and are both known to be potently transforming oncogenes(Kang, Bader et al. 2005, Liu and Roberts 2006, Huang, Mandelker et al. 2007). It stands to reason that if APOBEC-drives *PIK3CA* helical hotspot but not kinase hotspot mutations, helical hotspot mutations should be found in tumours with high APOBEC activity at a higher frequency than kinase-hotspot mutations. In order to test this, I examined *PIK3CA* mutations in a dataset of 10 different tumour types (See Methods for details). Using TCW -> TKW fraction estimates and a binomial GLM regressing the TCW -> TKW fraction against mutation hotspot and controlling for tumour type revealed enrichment for APOBEC mediated mutagenesis in helical-hotspot mutated tumours (OR = 1.6, $p < 0.001$).

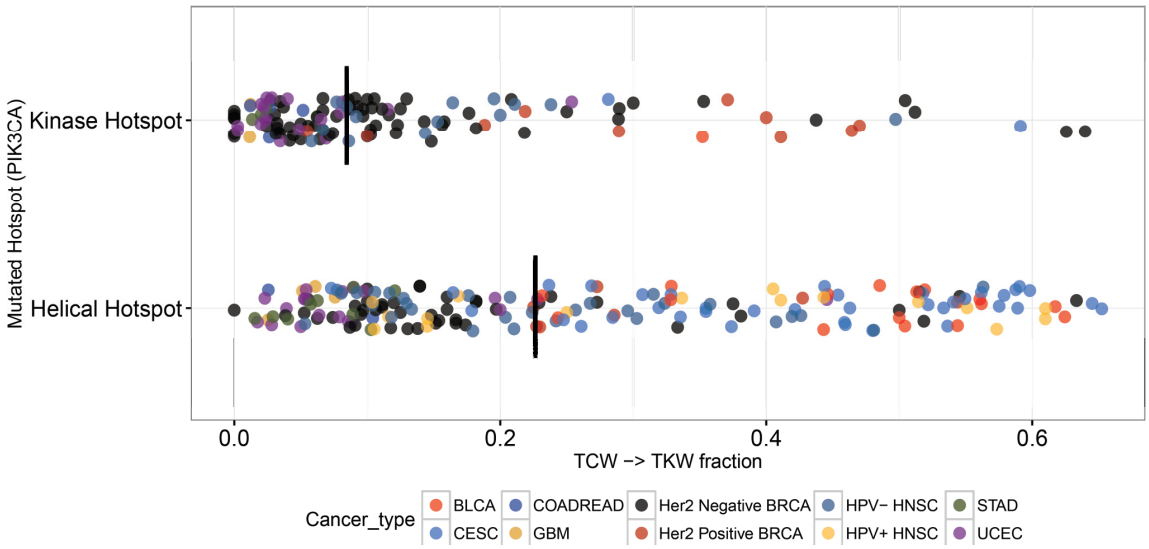


Figure 22: Breakdown of TCW -> TKW fractions across multiple tumour types with PIK3CA hotspot mutations. Y axis = mutated hotspot, X axis = TCW -> TKW fractions. Black crossbars represent groupwise medians and colours represent tumour type. BLCA = Bladder Cancers, COADREAD = Colorectal, BRCA = Breast, HNSC = Head and Neck Squamous Cell Carcinoma, STAD = Stomach Adenocarcinoma, UCEC = Uterine Corpus Endometrial Carcinoma, GBM = Glioblastoma Multiforme, CESC = Cervical Squamous Cell and Endocervical Adenocarcinoma.

This association was also significant using an independence test comparing TCW -> TKW fraction distributions **(Figure 22)** between Helical-hotspot mutant and Kinase-hotspot mutant tumours (p = 0.001). These findings suggest that, independent of selection pressure, mutational processes can play a significant role in determining the spectrum of mutations that occur in a tumour and an interesting question for future work is to see how much mutational spectra can be delineated into the output of selective pressures acting on pre-existing pools of mutations which are to some level determined by mutational processes.

Analysis of putative factors influencing APOBEC mediated mutagenesis.

Previous studies have implicated APOBEC3B expression at the mRNA level as a determinant of APOBEC mediated mutagenesis, especially in breast cancer (Burns, Lackey et al. 2013) (Burns, Temiz et al. 2013). Given the differences between HPV+ HNSC and HPV- HNSC in APOBEC-mediated mutagenesis, I tested for, and found significant differences in the expression of most members of the *APOBEC/APOBEC3* family (FDR < 0.001 , **Figure 23A**) . However, in agreement with previous studies on cervical cancer (Ojesina, Lichtenstein et al. 2014), only weak, albeit statistically significant, correlations were found with all members of the *APOBEC3* family **(Figure 23B)**, suggesting that breast cancer is exceptional (Roberts, Lawrence et al. 2013). Viral gene transcript abundance and TCW -> TKW fraction were not correlated either.

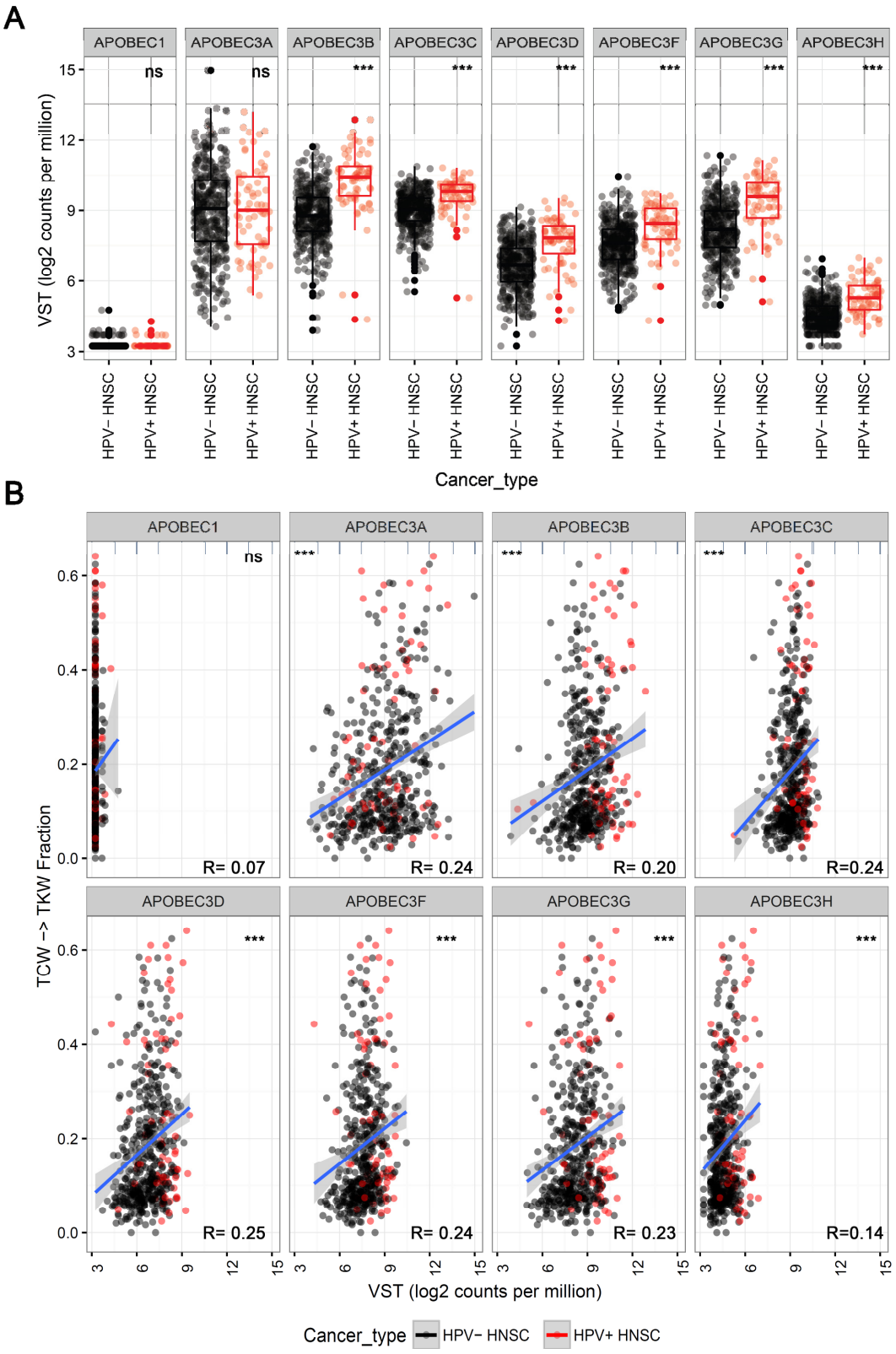


Figure 23: A) mRNA expression of APOBEC family genes in HPV+ vs HPV- HNSCs, * = FDR (BH) < 0.0001, Wilcoxon's Rank Sum Test. B) Correlation between TCW -> TKW fractions and mRNA expression of APOBEC family members. *** = FDR (BH) < 0.001, R = Spearman's Rho.**

These findings however come with the caveat that current expression levels of both viral and APOBEC transcripts may not correspond to levels of APOBEC activity when these mutations were actually generated in the course of tumour evolution.

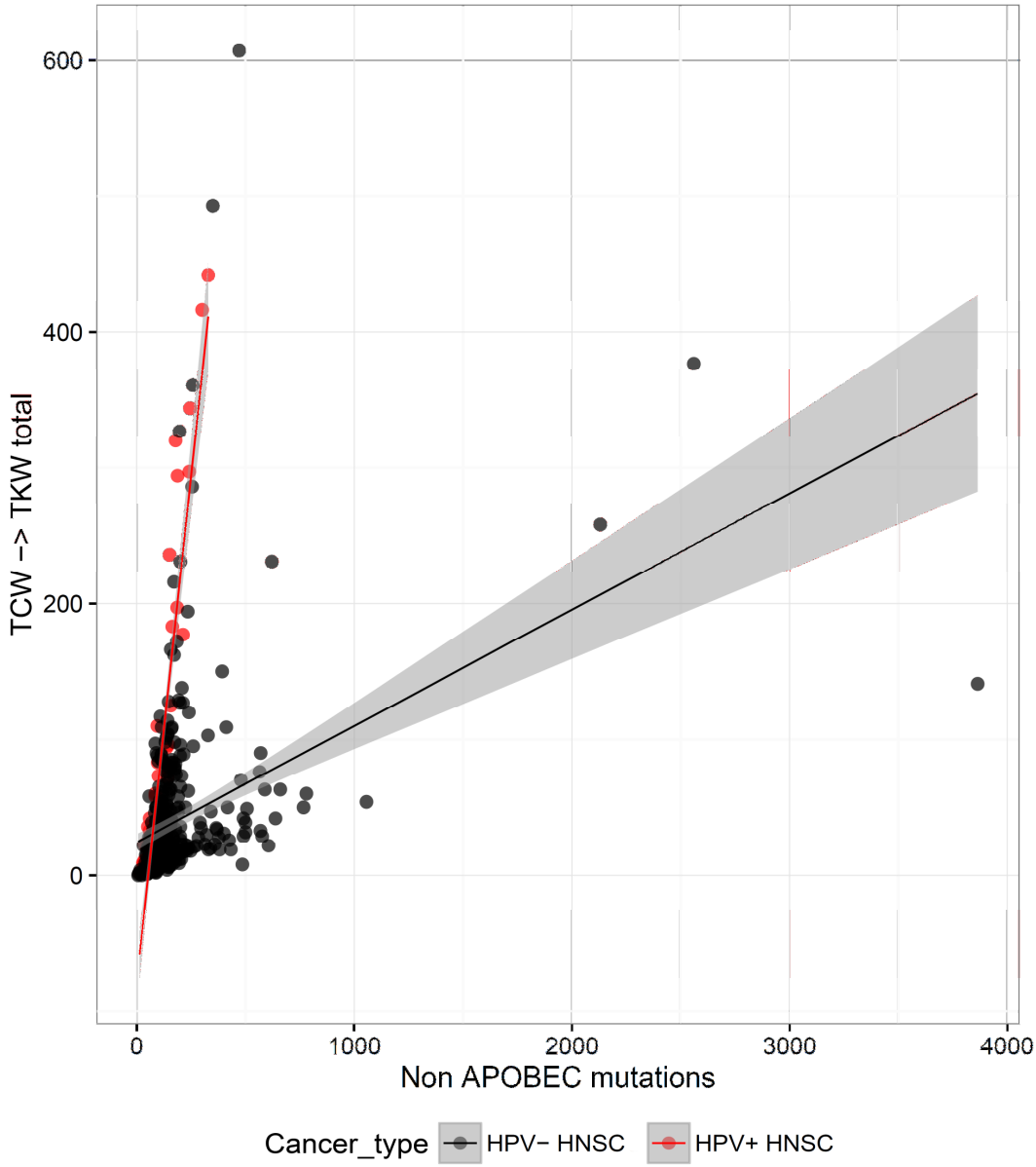


Figure 24: Relationship between overall mutational load and APOBEC mutation load stratified by HPV status. For a given tumour type there is a conserved relationship between the APOBEC-induced and general mutational loads. Grey areas represent 95% Confidence Intervals for linear model fits.

A striking linear association was however found between the APOBEC and non-APOBEC mutational burdens in tumours **(Figure 24)**, with the slope varying by tumour HPV status. In order to investigate the statistical relationship between tumour subtype, TCW -> TKW mutation ANOVA modelling was carried out. Model selection revealed the most variation (adjusted $R^2 = 0.37$) in the TCW -> TKW mutation total was expained by HPV status and the non-APOBEC mutational load interacting (both explanatory variables significant at $p < 0.01$, model F statistic significant at $p < 2.2e-16$).It has been recently shown that the repair of mismatches is associated with the generation of APOBEC-induced mutations (Chen, Miller et al. 2014) , as is the process of break-induced replication that generates ssDNA that serves as a substrate for cytosine deamination (Sakofsky, Roberts et al. 2014) and these processes offer potential explanations for the relationships observed between TCW and non TCW mutations (NB – this approach was originally devised for the interim analysis published in (Henderson, Chakravarthy et al. 2014) by Stephen Henderson).

Chapter Conclusions.

Until recently, it had been unclear if HPV, following the constitutive inactivation of E6 and E7 and accompanying transcriptional and epigenetic changes it could directly induce, was responsible for shaping the evolution and selection of subsequent genetic hits required for full transformation or just generated a population of cells with dysregulated p53 and pRb function that could then acquire additional hits through mutational processes in which HPV directly played no part. The evidence presented in this chapter suggests that APOBEC-mediated cytosine deamination is a connecting link between anti-HPV antiviral responses and the evolution of these additional mutations.

In HPV+ HNSC, which is traditionally associated with non-smokers and younger patients, there appears to be greater dependence on APOBEC to generate the necessary mutations for malignant transformation, and there are features of the transcriptional signature associated with HPV-driven tumourigenesis that potentiate APOBEC activity.

While the correlation between *APOBEC* expression and activity was found to be weak, it is important to realise that expression levels at sampling may not correspond to expression levels when the mutations in these genomes were generated, which may have occurred across a wide range of time-points in tumour evolution.

Chapter 5: Common Epigenetic Profiles Unify HPV-driven cancers across tissues.

The relationship between various datasets used for signature discovery, training and validation are summarised in a flowchart presented as part of Appendix A3.

Statistical analyses establish a comprehensive catalogue of DNA methylation changes in HPV-driven tumours and point towards a hypermethylator phenotype.

Using a discovery set of 844 samples (see methods for class/tissue/HPV detection method breakdown) I defined signatures of methylation variable positions (MVPs) and Differentially Methylated Regions (DMRs) for HPV-driven tumours that could distinguish them from HPV- tumours and normal controls that arose in similar anatomical sites. Using a limma based pipeline for MVP-calling, I identified a total of 8225 MVPs differentially methylated with a beta-value median shift of at least 20% at an FDR (Two-step BH adjusted) of 0.001 or less **(Figure 25)**. The signature was robust to tissue type, with tissue adjusted analyses yielding 6870 of the same MVPs at the same thresholds.

Of these MVPs, 6753 were hypermethylated in HPV+ cancers, and only 1472 were hypomethylated, pointing towards a hypermethylator epigenotype ($p < 2.2e\text{-}16$, 95% CI = 0.81 – 0.82 towards hypermethylation, binomial test against null probability of no skew). In addition to MVPs, analysis was also conducted using fDMR. This approach clusters probes by annotation based on number of statistically significant MVPs, and functional annotation. It then accounts for the spatial correlation of the probes in the candidate DMR and performs Stouffer-Liptak p.value combination to identify statistically significant DMRs. 781 DMRs, mapping to 3328 and 508 hypermethylated and hypomethylated MVPs respectively, were discovered to be associated with HPV status, with each DMR containing at least 3 MVPs (defined as described above) at DMR FDR < 0.001 and a DMR-median shift of 20% in the beta-value.

Of these, 664DMRs were hypermethylated and 117 were hypomethylated in HPV+ samples, again supporting a hypermethylator epigenotype at the DMR level (CI = 0.82 – 0.87, p < 2.2e-16,binomial test).

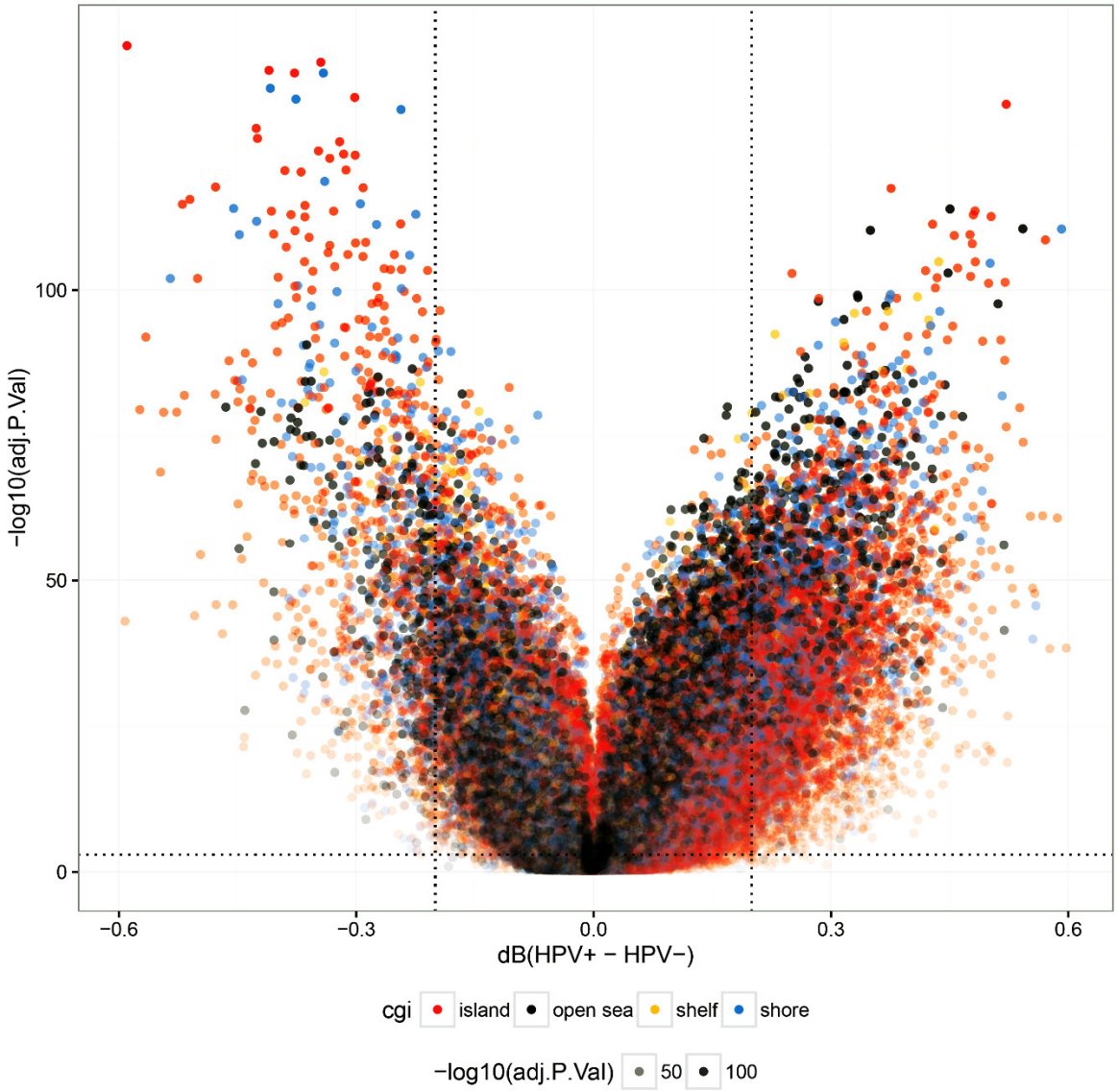


Figure 25: Volcano Plot showing differences between HPV+ and HPV- samples. X axis = Beta value difference (dB > 0 when hypermethylated in HPV+ tumours). Y axis = -log10 (False Discovery Rate). Colours indicate relationship between probes and CpG Island related annotations.

The Global Hypermethylator Phenotype Extends to Most Categories of HM450k probes.

Having established the initial methylation signature for HPV-driven tumourigenesis and having discovered a global trend towards hypermethylation, I examined skews towards hypermethylation across CpG-density and transcript-centric categories of probes to decipher patterns of global methylation in these cancers.

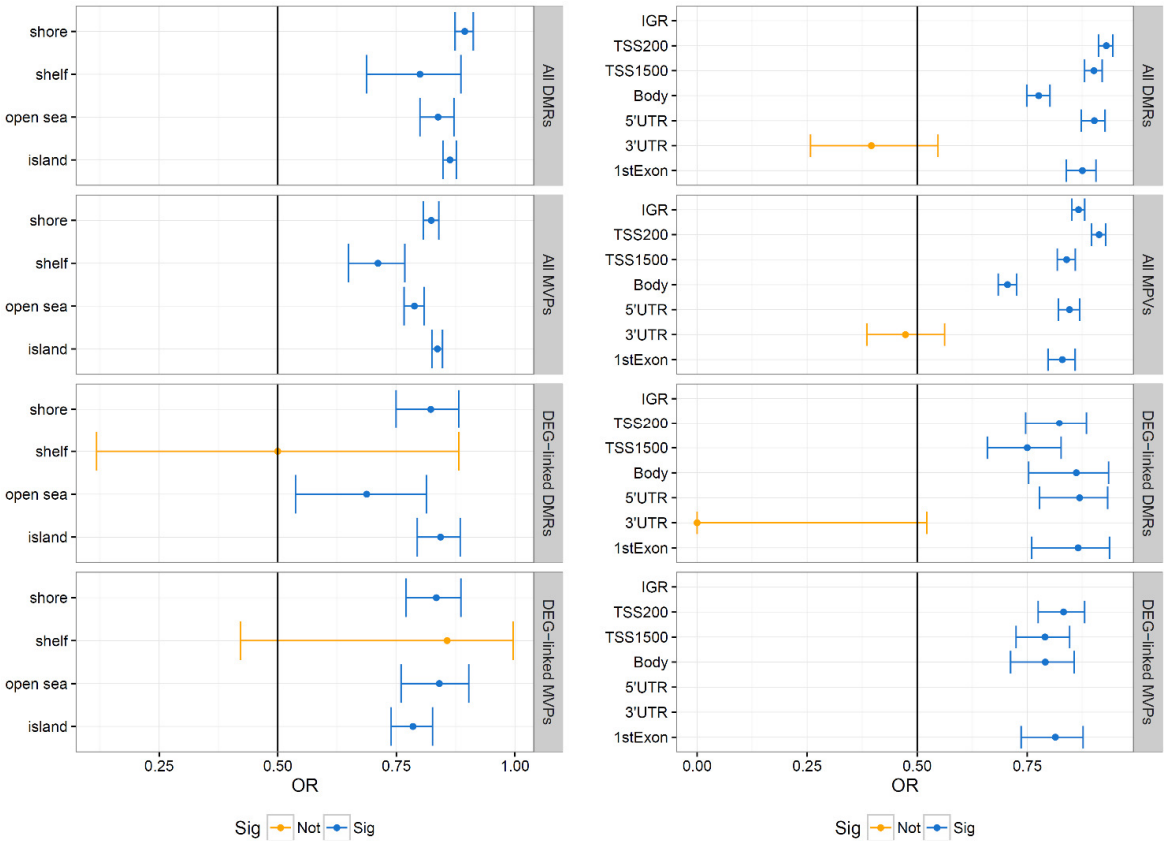


Figure 26: Numbers of hypermethylated vs hypomethylated probes by probe and signature. X axis = 95% CI of probability of hypermethylation , with vertical line representing equiprobability of hypermethylation and hypomethylation

Apart from CpG shelf associated probes that mapped to the DEG-linked MVPs, and 3'UTR probes that mapped to DMRs and all MVPs, and gene body probes that mapped to DEG-linked DMRs, all other categories demonstrated a distinct skew towards DNA hypermethylation in HPV+ tumours (FDR < 0.05) **(Figure 26)**.

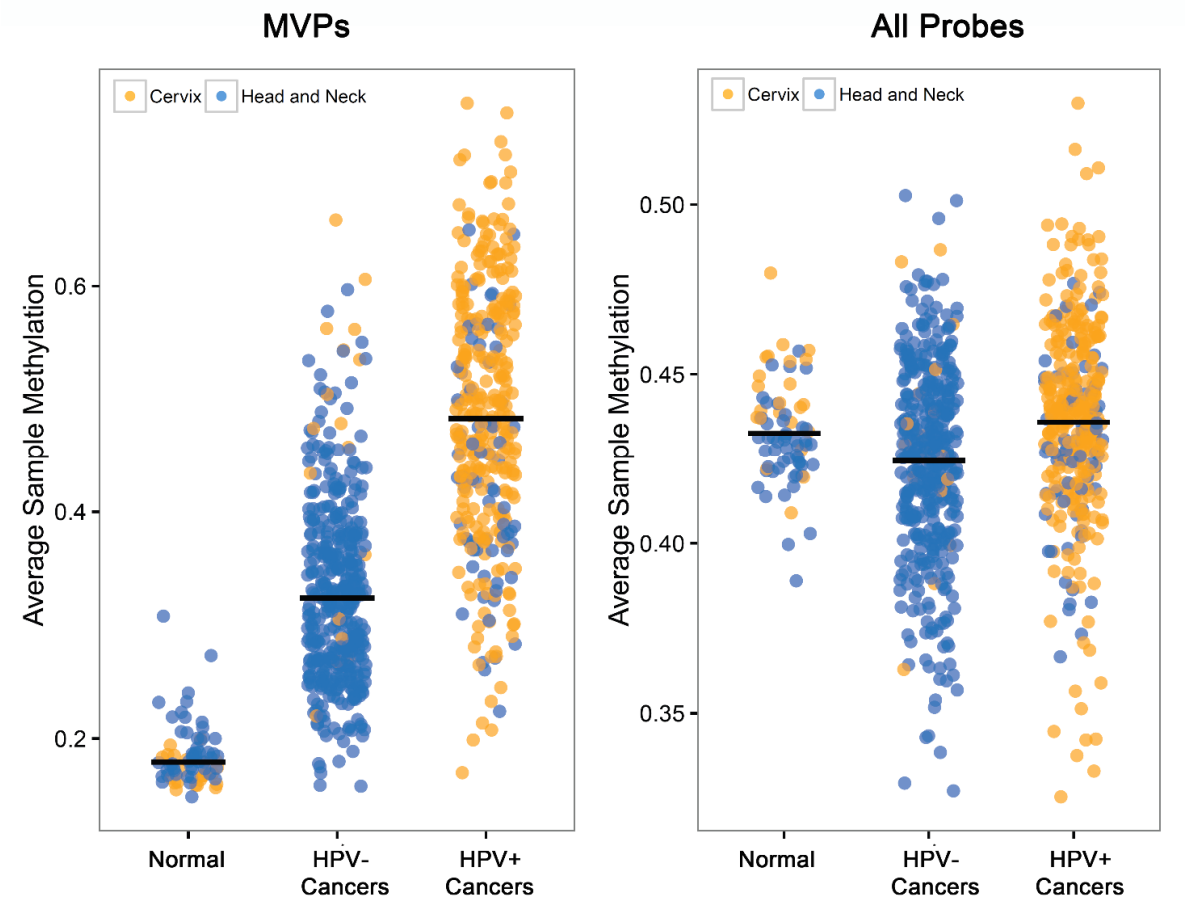


Figure 27: Distributions of average beta-value in differentially methylated probes and all 450k probes respectively. Y axis = average methylation value per sample, colour represents Tissue type, crossbars represent group medians.

While previous research has reported HPV-associated hypermethylation, it has been based on comparisons of HPV+ and HPV- cancers with no normal samples available (Lechner, Fenton et al. 2013). These patterns could in fact have been produced by hypomethylation in HPV- tumours rather than hypermethylation in HPV+ tumours, relative to normal tissue. However, examining the distributions of per-sample means for MVP associated probes in normal epithelial samples (Cervix/Head and Neck), HPV- cancer and HPV+ cancer showed marked hypermethylation in the latter relative to the former two groups ($p < 2.2e-16$) **(Figure 27)**. When considering array-wide distributions however, differences were not significant between these categories at FDR < 0.05 and shifts in distributions were minor. The attribution of a hypermethylator phenotype to HPV+ tumours relative to normal tissue and HPV- tumours therefore remains dependent on whether comparisons are made within differentially methylated probes or universally.

Pan-tissue epigenetic signatures are useful for classification by HPV status.

Having derived methylation signatures at the probe and region levels for HPV-driven tumourigenesis, I examined if the features contained therein could serve as a basis for accurate classification of HPV-status by developing Random Forest classifiers and visualising clustering patterns of feature methylation values. When the discovery set was initially visualised on heatmaps **(Figure 28)**, the probes in these signatures stratified samples into distinct clusters that track with HPV status. Evaluation of these features for classification using Random Forest classifiers produced highly reliable classifiers.

The MVP signature returned Positive (PPV) and Negative (NPV) Predictive Values (See Methods) of 0.96, while the DMR signature returned values of 0.95 and 0.94 when based on out-of-fold predictions, serving to internally validate these signatures.

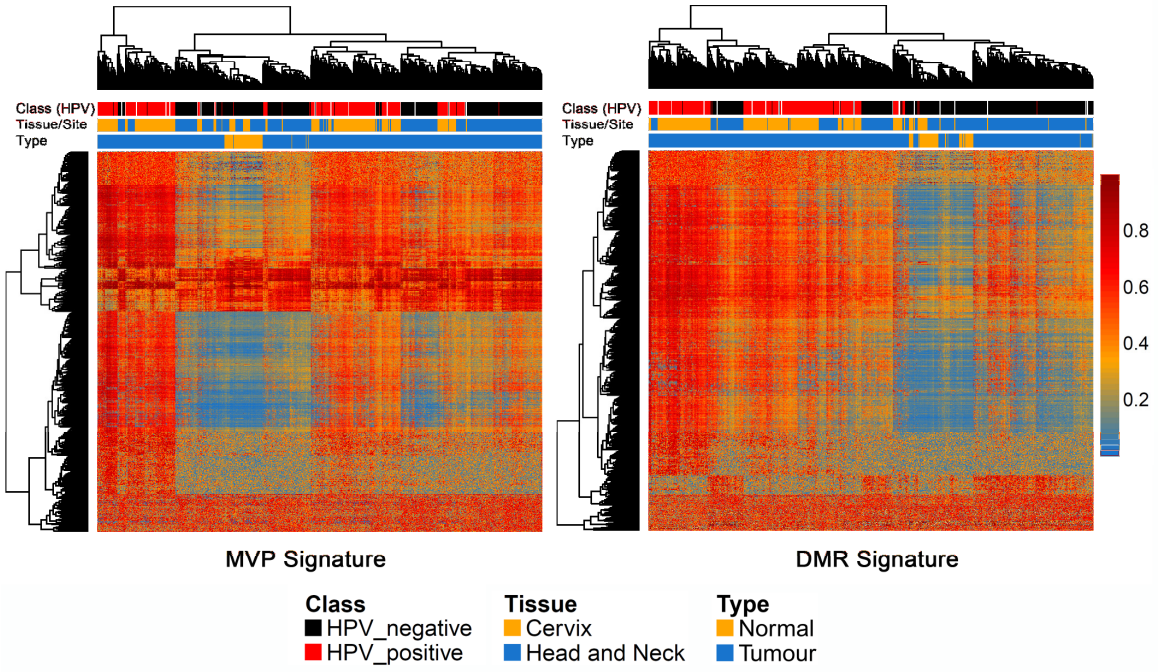


Figure 28: Pan-tissue methylation signatures for HPV-driven tumourigenesis group HPV+ tumours together in distinct blocks. Columns represent samples and rows represent probes. Intensity represents beta value for methylation.

The classifier was then applied to independent datasets to test concordance with HPV status as derived by other means. In a small dataset of Penile Carcinomas (n=26, HPV+ = 7) (PeCa) **(Figure 29A)**, where HPV status was only known on the basis of p16 staining and DNA in-situ hybridization, PPV of 0.83 and NPV of 0.89 (DMR), and PPV of 1.00 NPV of 0.9 (MVP) were observed. It is possible that some of the misclassified samples were not truly HPV+ given some samples had HPV at less than 1 copy/cell, and evidence for active HPV transcription was lacking.

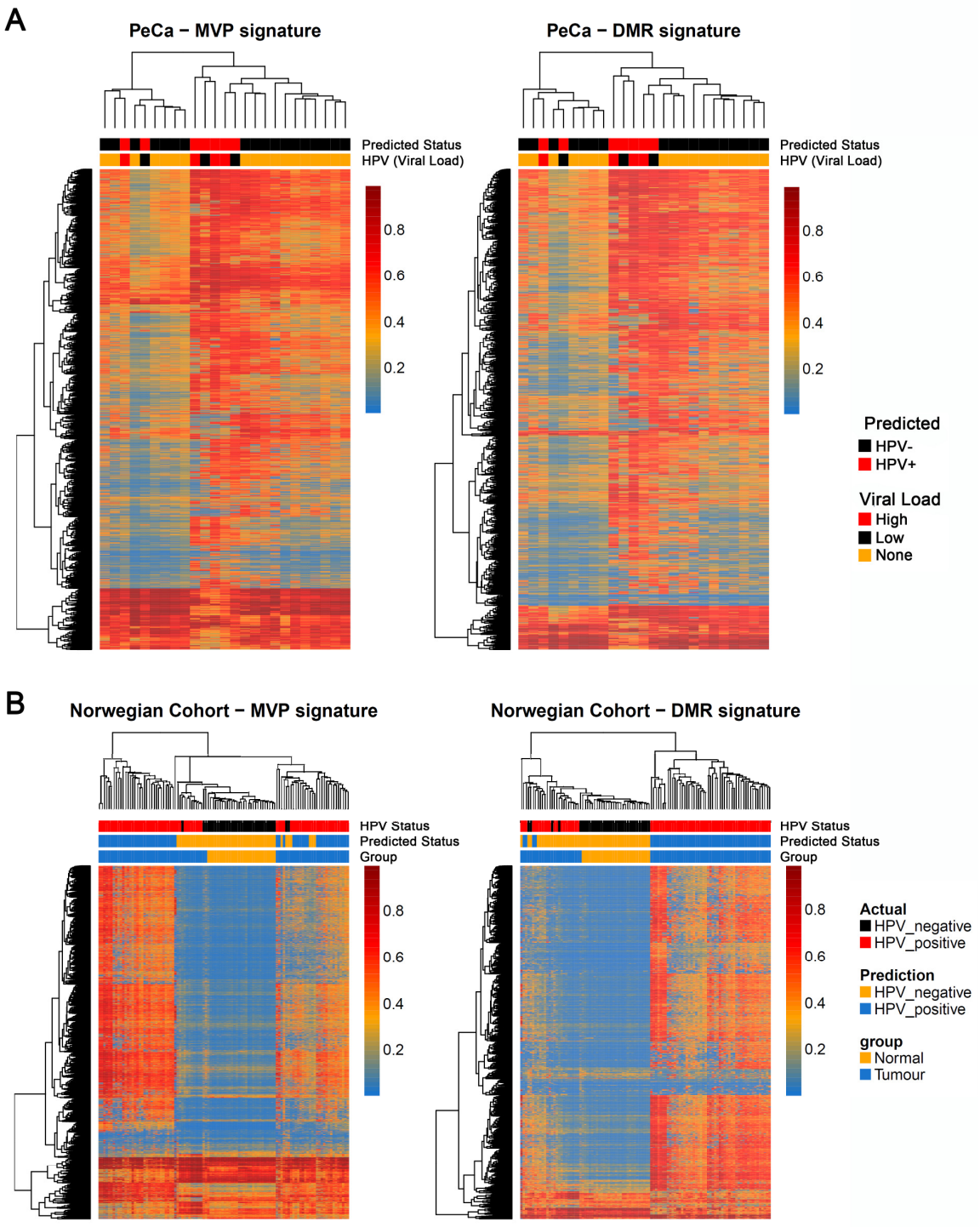


Figure 29: (A) Patterns of MVP and DMR signature probe methylation in Penile Carcinomas. (B) MVP and DMR signature methylation in an independent cohort of No rwegian Cervical Cancers. Intensities represent beta-values, for PeCa, annotations represent HPV status prediction and HPV viral load classification. For Cervical Cancers, annotations represent HPV status by RNA-seq (Normal samples assumed HPV-) , Random Forest predictions and Cancer/Normal status.

Application to an independent Cervical Cancer dataset from Norway, with samples obtained from our collaborator Helga Salvesen and samples processed by Dr Andrew Feber and UCL Genomics, (29 Normal, 77 tumour) suggested that the vast majority of cancers recapitulated methylation patterns seen in the discovery set **(Figure 29B)**. However, there were multiple samples that were misclassified, resulting in high PPV (1.0) and low NPV (MVP = 0.69, DMR = 0.66) that were HPV+ by the detection of viral transcripts, despite clustering patterns clearly separating HPV+ and HPV- samples.

This indicated the occurrence of overfitting or between-batch technical variation that the normalisation procedure could not have accounted for. These explanations were tested by first splitting the discovery set into two, training on one, and then using the other half as an independent test sets. For both the MVP and the DMR signatures, internal cross-validation accuracy was better than independent validation using the other half (Kappa = 0.86 vs 0.91 for DMRs, 0.88 vs 0.93 for MVPs).

Models trained on half the discovery set also demonstrated better performance in the Norwegian cohort compared to those trained on the entirety of the discovery cohort; (NPV for MVPs = 0.79, DMRs = 0.76, vs <0.7 for full model) implicating overfitting as the reason for worse performance in independent datasets. Potential technical variation between datasets was also identified upon visualising an MDS plot of MVPs **(Figure 30)**. Finally, estimating model performance using cross-validation while training a Random Forest on the Norwegian cohort using the signatures defined with the discovery set again demonstrated markedly higher accuracy (For MVPs, Kappa = 0.93, PPV = 0.91, NPV = 1. For DMRs Kappa = 0.86, PPV = 0.88, NPV = 0.97), confirming poor performance was not motivated by poor feature selection.

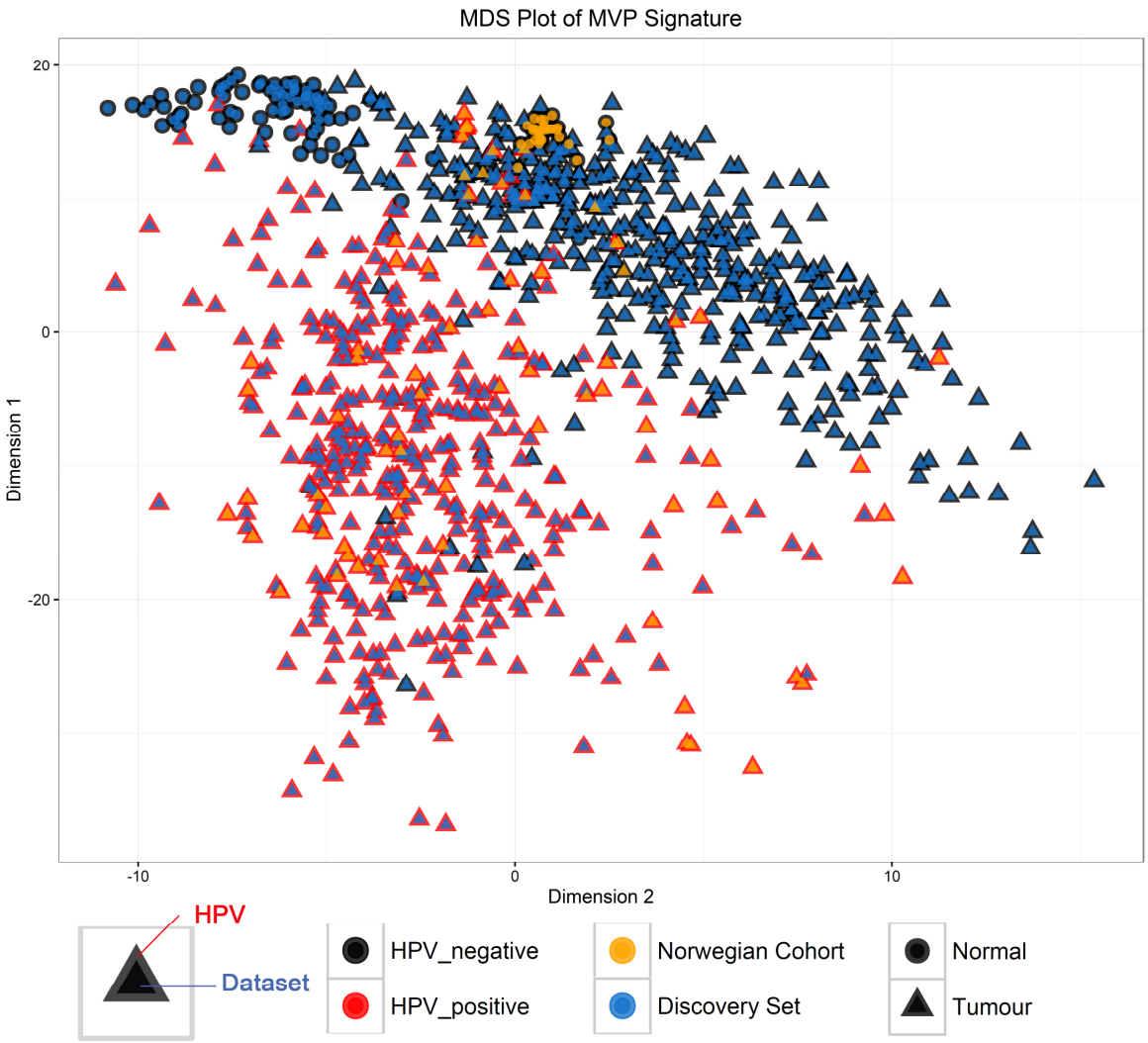


Figure 30: MDS plot of signature MVPs in the discovery and Norwegian cohorts, which were both normalised separately using Functional Normalisation. Some samples cluster by dataset and display dataset specific relationships between groups defined by HPV status, hinting at batch effects.

These findings, taken together, highlight the utility of pan-tissue DNA methylation profiles as a marker for HPV-driven tumourigenesis, highlight the threat of overfitting, and finally point to an unmet need for single-sample normalisation procedures to deal with technical variation on the 450k platform.

Integrating methylation with matched expression data identifies novel HPV-associated transcriptional changes.

The discovery dataset in large part was derived from samples processed by the Cancer Genome Atlas, ensuring the availability of matched RNA-seq samples for investigating gene expression. In total, out of the 844 samples in the curated Pan-tissue dataset, 794 had matched RNA-seq data.

Limma-trend analysis was used to identify genes that were differentially expressed and methylated as canonically expected for DNA methylation (overexpressed with increasing gene body methylation, and underexpressed with increasing promoter/1st Exon /5'UTR methylation).

In addition to these candidates, additional annotation was also performed to uncover long-range changes in expression regulated by enhancer methylation. Methylation at distal regulatory regions has been shown to be better correlated with gene expression changes than promoter methylation, and large profiling projects such as the FANTOM5 consortium have used CAGE-seq to sensitively document a wide array of enhancers that are marked by bidirectional transcription in a broad set of samples, along with establishing putative correlated gene pairs across tissues. In order to identify such distal methylation changes, the set of MVPs was reduced to those overlapping known FANTOM5 enhancers. 411/530 of these were found to be actively transcribed in at least one out of a collection of 11 samples of normal and cancer tissue types profiled in the discovery set at more than 0.5 tags/million.

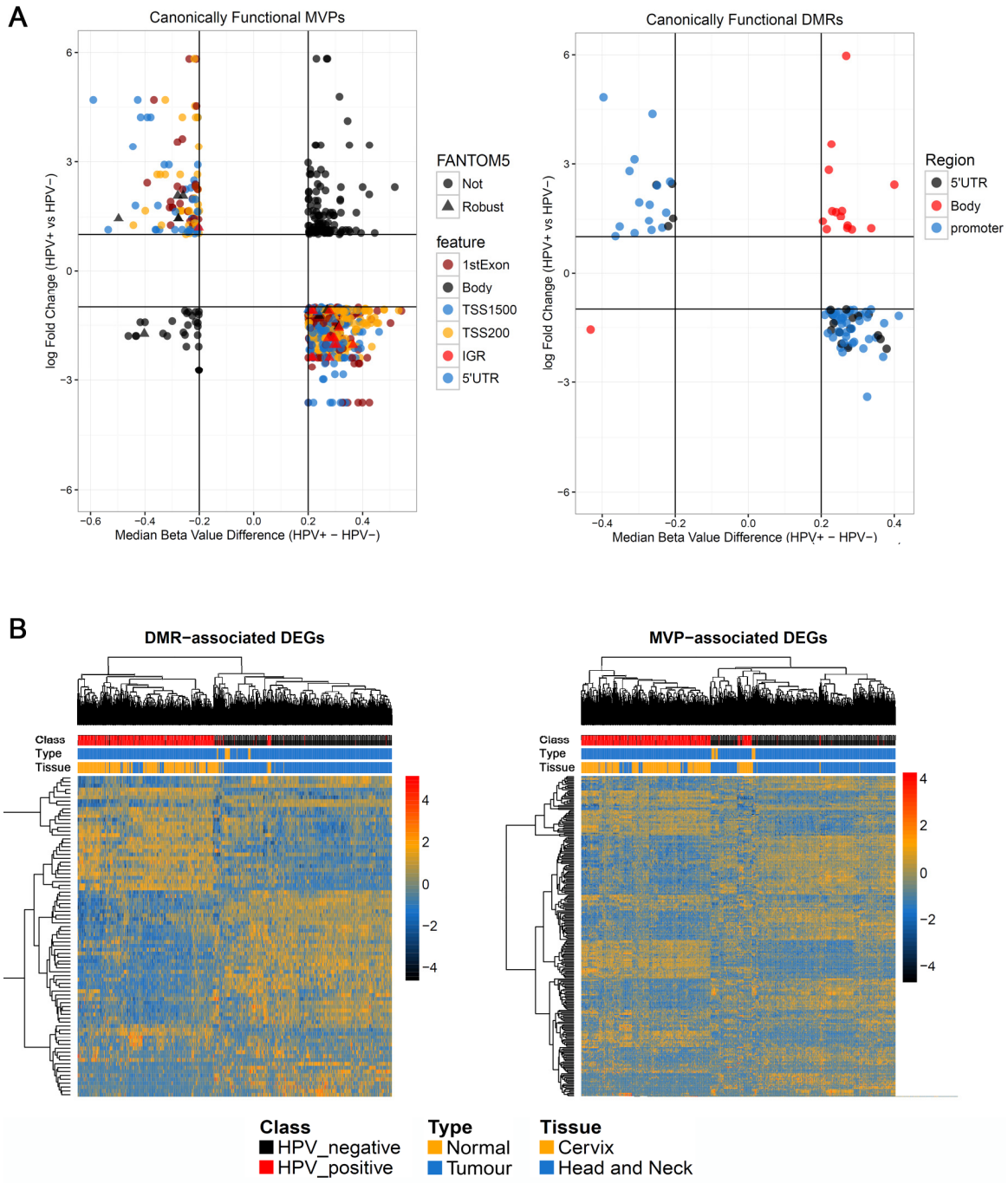


Figure 31: A) Quadrant plots showing genes that overlap differentially methylated probes/regions and are differentially expressed. B) Heatmaps of gene expression in 794 samples that overlap the methylation signature discovery set. Methylation associated Differentially Expressed genes accurately cluster HPV+ and HPV- samples.

These were then checked for inverse differential expression of correlated gene pairs. Of these, ultimately, 23 enhancer probes were differentially methylated, and were linked to canonically differentially expressed genes. Out of all the probes in the MVP signature, 684 probes mapped to 243 DEGs while 93 DMRs were canonically associated with 84 DEGs (Quadrant-plot of Expression-Methylation relationships in **Figure 31A)**. The functional MVP and DMR signatures, and associated DEGs, result in accurate clustering of samples by HPV status when visualized on heatmaps **(Figure 31B)**. The low proportion of DNA methylation changes associated with gene expression changes is not surprising as cancer cells with mostly ablated methyltransferase activity fail to derepress genes on a large scale (Blattler, Yao et al. 2014). Previous studies involving HPV-driven cancer methylomes, including the analysis of HNSC methylomes by (Lechner, Fenton et al. 2013), Penile Cancer methylomes (Feber, Arya et al. 2015) and Cervical cancers by (Farkas, Milutin-Gasperov et al. 2013) have all been constrained by the non-availability of matched expression data. The sets of epigenetically regulated transcriptional events derived here offer refined understanding of how epigenetic dysregulation shapes transcriptional events in HPV driven cancers.

Pathway Analysis contextualises the contribution of DNA methylation to transcriptional dysregulation in HPV+ tumours.

Efforts to characterize the functional context of canonically regulated Differentially Expressed genes was initially carried out using IPA for both DMR-associated and MVP-associated DEGs. No significant pathways or upstream regulators were discovered for DMR-associated DEGs.

Pathway analysis of MVP-associated genes however uncovered interesting sets of pathways dysregulated by DNA methylation. Two canonical pathways were found to be enriched at BH-FDR < 0.05; Embryonic Stem Cell Differentiation into Cardiac Lineage and Transcriptional Regulatory Network in Embryonic Stem Cells.

DMRs and MVPs not mapping to distinct pathways have two implications – either epigenetic events may simply contribute to transcriptional remodeling by dysregulating specific nodes in networks and pathways to which other mechanisms of dysregulation also contribute, or they simply comprise a passenger phenomenon. If the former hypothesis is true, it follows individual DMRs or MVPs should map to genes that are either part of global transcriptional networks associated with HPV+ tumourigenesis or have gene functions directly relevant to HPV-driven tumourigenesis. Correlation analysis and testing for overlaps with either the HPV-metasignature or literature searches for associations with HPV-driven tumourigenesis were then used to nominate such genes.

Three DMRs were found to be associated with the HPV-metasignature, and displayed differential expression by HPV status and a strong correlation between methylation and expression. *SYCP2* (R=-0.75), *HLTF* (R=-0.60) and *RPA2* (R=-0.58, all Spearman's Rank Correlation) and were hypomethylated in HPV+ tumours **(Figure 32A)**. Amongst the 243 canonical MVP-associated genes, 10 genes were found to overlap with the HPV-metasignature ($p = 0.0001$ for enrichment, randomisation test) . These findings together suggest epigenetic alterations contribute significantly towards overall transcriptional dysregulation.

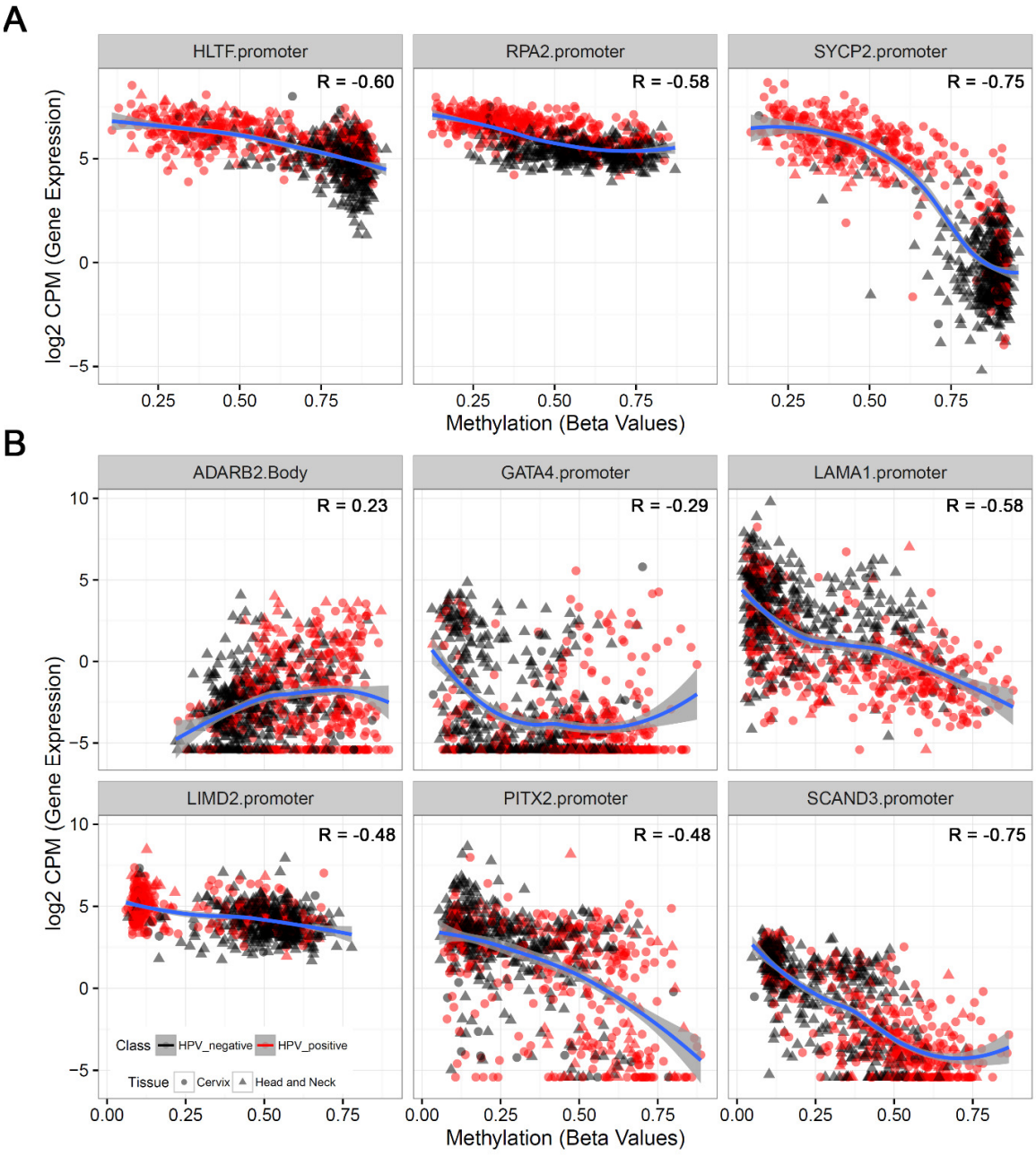


Figure 32: A) DMRs overlapping with the HPV-metasignature B) Novel DMRs are highly correlated with gene expression and show divergent transcriptional patterns by HPV status.

In addition to the epigenetic changes described above, several DMRs are strongly correlated with expression and have either known roles in HPV-driven tumourigenesis or are unexplored and novel candidates **(Figure 32B)**.

PITX2 is a promoter that is highly methylated in HPV+ tumours relative to HPV- tumours, and is a candidate for a driver epigenetic event because PITX2a has previously been shown to actively interfere with HPV18 E6-mediated degradation of p53 by binding to the former (Wei 2005), disrupting an interaction that is central to HPV-driven tumourigenesis.

RPA2, which is also dysregulated in the HPV-metasignature, is a regulator of mitotic catastrophe and maintains replication fork integrity after stalling due to replication stress (Murphy, Fitzgerald et al. 2014).

Novel candidates include *SCAND3,* which is not only associated with a DMR and is strongly correlated, but is also associated with validated long-range enhancers. *SCAND3* encodes a member of a family of retrotransposon-derived transcriptional regulators (Llorens, Bernet et al. 2012) and the functional significance of this gene is unclear.

Other novel candidates regulated epigenetically include *GATA4*, which encodes a transcription factor that is known to be a tumour suppressor in malignant astrocytomas (Agnihotri, Wolf et al. 2011), LIMD2, a known prometastatic factor (Peng, Talebzadeh-Farrooji et al. 2014) and *ADARB2*, which encodes a member of the family of RNA Adenosine Deaminases.

A tale of two populations: Follow-up analysis on KRT7 methylation patterns suggests HPV-type associated tropism for distinct precursor cell populations.

Other DMRs may have implications in explaining the cellular origins of these HPV+ tumours. It has been postulated that cervical cancer almost exclusively arises from a squamocolumnar junction layer of epithelial cells with foetal origins (Herfs, Yamamoto et al. 2012) and one of the defining markers of this population is high levels of *KRT7* expression, which is marked in HPV+ tumours by a hypomethylated promoter-DMR associated with overexpression and strong correlation (R= -0.7, Spearman's Rank Correlation, **Figure 33A**) , suggesting that *KRT7* expression and hypomethylation may act as a cellular marker for the expansion of squamocolumnar junction cells. Interestingly, comparing distributions of DMR methylation and *KRT7* expression suggested differential expression between HPV+ Head and Neck Cancers and Cervical Cancers.

Visualising the distributions of *KRT7* expression and promoter DMR methylation in HPV+ HNSC (mostly HPV16+) and CESC positive for HPV16/HPV18/HPV45 demonstrated that the distributions of HPV16+ CESC and HPV+ HNSC largely overlapped, whereas patterns differed for HPV18+ and HPV45+ compared to HPV16+ cervical cancers or HNSC, with very high expression levels of KRT7 (and very low methylation levels) the norm in HPV45+ / HPV18+ samples **(Figure 33B)**. This led to detailed analyses of squamocolumnar junction marker genes in HPV+ tumours in this dataset.

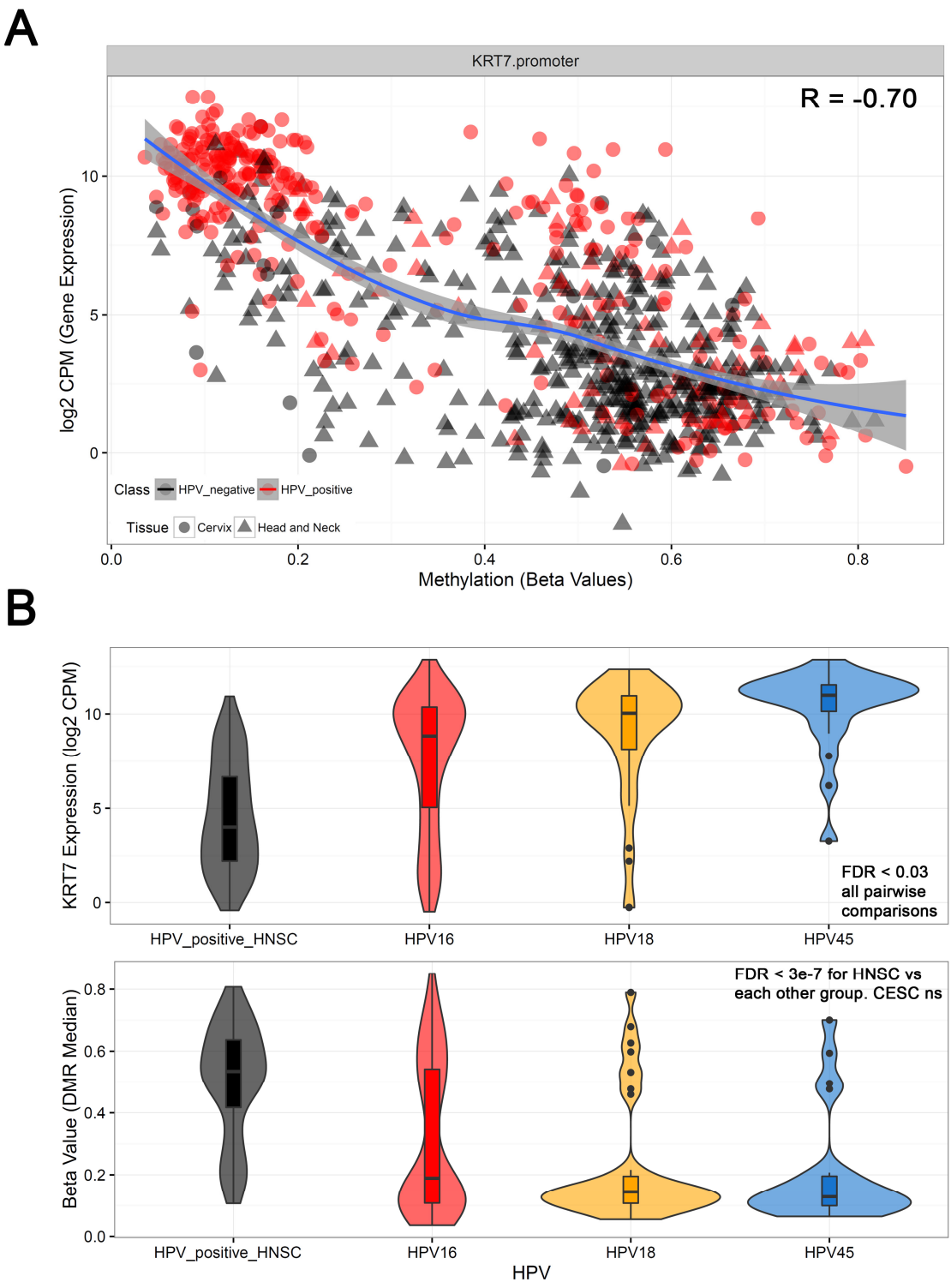


Figure 33: A) KRT7 Expression and promoter DMR methylation are strongly anti-correlated. B) Violin plots demonstrate variations in KRT7 Expression and Methylation by group. FDR from Holm correction on pairwise Wilcoxon's Rank Sum Tests. HPV16, HPV18 and HPV45 represent groups of cervical cancer samples with these HPV types. The small number of adenocarcinomas in the dataset makes confounding by cervical histology unlikely.

Signature gene lists for squamocolumnar junction cells were obtained relative to columnar and squamous epithelial cells from (Herfs, Yamamoto et al. 2012) (See Methods), and HPV+ HNSC/ HPV45+/HPV16+/HPV18+ CESC were Consensus clustered on this basis. Two clusters were identified as the most robust solution, and one cluster contained the vast majority of the HPV18+ and the HPV45+ tumours (**Figure 34)**.

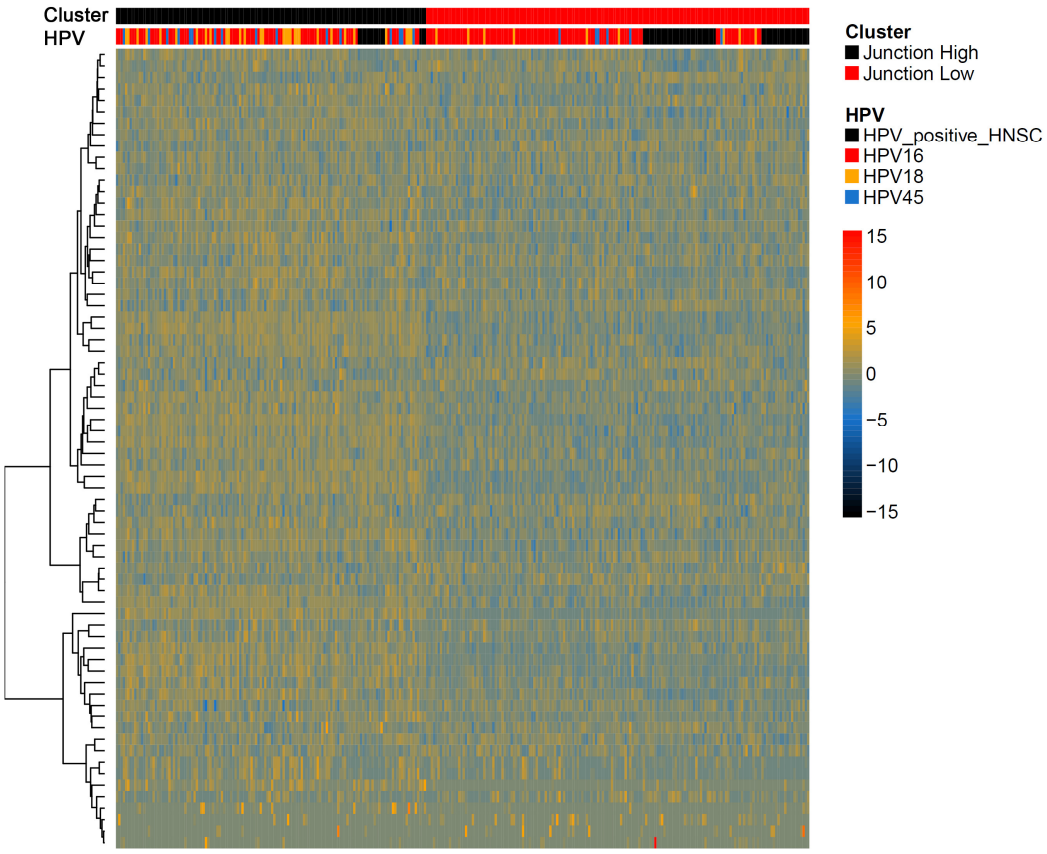


Figure 34: Consensus Clustering based on a squamocolumnar junction cell expression signature identifies that HPV16 may infect cells not marked by this signature, whereas HPV18 and HPV45 are more restricted.

85% of HPV+ HNSC and 57% of HPV16+ CESC mapped to the Junction-low signature, suggesting that while HPV16 could transform a broad range of cellular populations that differed in patterns of squamocolumnar junction gene expression, whereas HPV45 (73% Junction-High) and HPV18+ (77% Junction-High) tumours were more restricted to a cellular population defined by the high-expression cluster.

The implications of these findings are two-fold. Firstly, the abundance of the types of cells analogous to the Squamocolumnar Junction cells of the Cervix may determine what HPV types infect and transform what tissues and at what frequency, and secondly, different HPV types may be associated with tumours that display heterogeneous transcriptional and epigenetic signatures, and consequently behaviour when transformed. The latter possibility is comprehensively evaluated in the next chapter.

Chapter Conclusions

The work in this chapter contains the largest analysis of methylomes from HPV+ cancers and HPV- tumours that arise in similar tissues. A signature of global hypermethylation marks HPV+ tumours, even if only a small fraction of the genes differentially methylated between HPV+ and HPV- tissues display evidence for associated expression changes. DNA methylation appears to contribute to transcriptional dysregulation in concert with other mechanisms that alter transcription and does not act as a mediator of co-ordinated sets of transcriptional changes.

Despite this, putative candidate DMRs like those at *PITX2* likely regulate the expression of candidate driver epigenetic events, and one of the future challenges that may emerge in the analysis and understanding of methylation data is the development of methods that can infer driver methylation events; a task that has been made possible for mutations using multiple criteria.

Finally, DNA methylation, being amenable to be used as a biomarker, renders the signatures defined in this chapter a source for biomarkers that may serve to unambiguously identify HPV+ tumours in a clinical setting.

Chapter 6: Categories of Viral Classification Correlate with Clinical and Molecular Heterogeneity within HPV+ Cervical Cancers.

The main steps of analyses undertaken in this chapter have been described in a flowchart in Appendix A3.

Molecular heterogeneity is a function of taxonomic variation in Human Papillomaviruses.

Of the various cancers caused by HPV, HPV16 predominates in cancers of the head and neck, whereas in cervix HPV18 and HPV45 also play a significant role, with minor contributions from other HPV types (Saraiya, Unger et al. 2015). The overrepresentation of HPV16 across all HPV-driven cancers raises questions of whether there are critical differences between HPV16-driven tumours and other HPV driven tumours in terms of the molecular profiles they establish, the cells of origin for the tumours they cause, and given the ability of HPV16 to induce the pan-tissue signatures described earlier in this thesis, and finally if differences in viral genomes translate to different clinical behaviour through establishment of varied molecular profiles.

HPV18+ cancers have been associated with poorer outcomes following radical hysterectomies in early stage cervical cancer (Burger, Monk et al. 1996), likely as a consequence of greater extents of invasion and nodal metastasis (Im, Wilczynski et al. 2003). More recently, it was shown that a small subset of 30 CpGs (referred to as F30 hereon) sites classified cervical cancers into two groups with different prognostic outlooks, with one of the groups predominantly containing HPV16+ cancers (Feber, Arya et al. 2015). In order to investigate the relationship between clinical outcomes and taxonomic categories of HPV, and to catalogue the molecular heterogeneity mediating these associations, RNA-seq was used to allocate HPV type to a total of 307 tumours in total, followed by analysis of transcriptome and epigenomic variation (HPV status calling by Stephen Henderson and from (Tang, Alaei-Mahabadi et al. 2013)).

HPV-types were then allocated to their respective clades and restricted to A7 and A9, comprising 291 tumours and into the F30 clusters. The original F30 publication only examined a subset of cervical cancers, so I initially checked and confirmed that F30 clusters showed associations with Clade (89.1 % of A7, and 16% of A9, were negative for the good survival F30 signature, OR=5.5, p = 1.6e-12, Fisher's exact test) and Overall Survival (HR= 0.38 if F30+, p = 0.001), matching expectations on both counts.

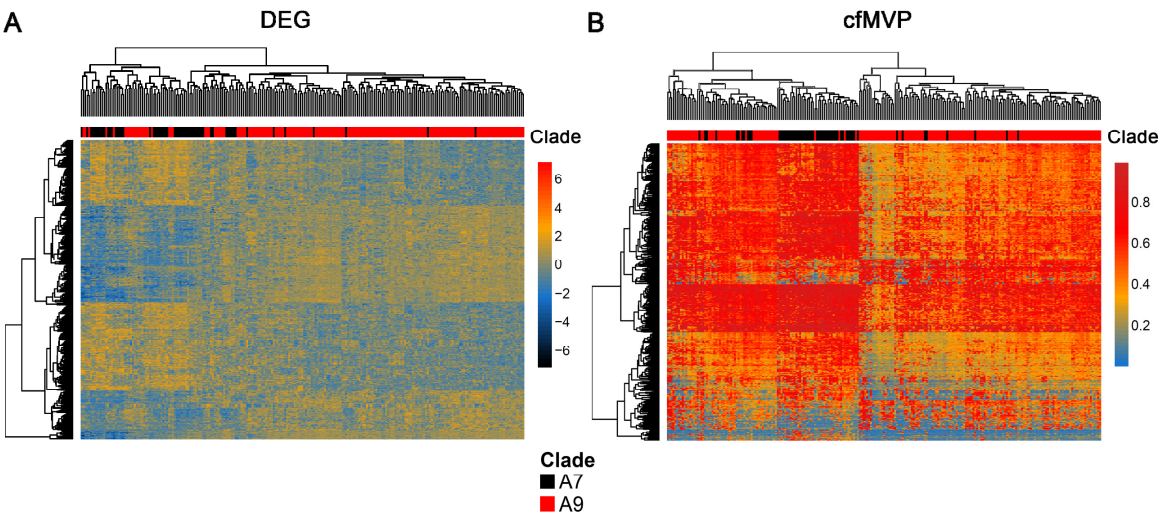


Figure 35: Differentially Expressed Genes (2FC, FDR < 0.01) (A) and canonically functional MVPs (delta Beta > 0.1, FDR < 0.01) (B) from comparing Clade A7 with A9. Annotation Ribbon indicates clade.

Having established that HPV clade was associated with a molecular pattern of prognostic utility, comparative analyses of methylomes and expression signatures between Cervical Squamous Cell Cancers (n = 233) driven by HPV types in Clades A7 and A9 were carried out, and documented 596 differentially expressed genes (2FC, FDR < 0.01) and 592 MVPs (mean beta-value difference 0.1, FDR < 0.01) that were canonically associated with 252 differentially expressed genes (2FC, FDR < 0.01) **(Figure 35)**.

Comparing survival between clades revealed differences in outcomes (HR = 0.512, p = 0.045, LRT p value = 0.05) overall and a bigger difference when comparing clades within early stage (Stage I and Stage II) tumours (HR = 0.39, p = 0.018, LRT p value = 0.0242). These initial analyses suggested that early stage tumours showed more marked clade associated clinical differences and established that HPV clades associate with molecularly distinct categories, in the process motivating the use of early stage tumours as a dataset to study molecular differences in tumours as a function of taxonomic classification of the HPV-types driving them.

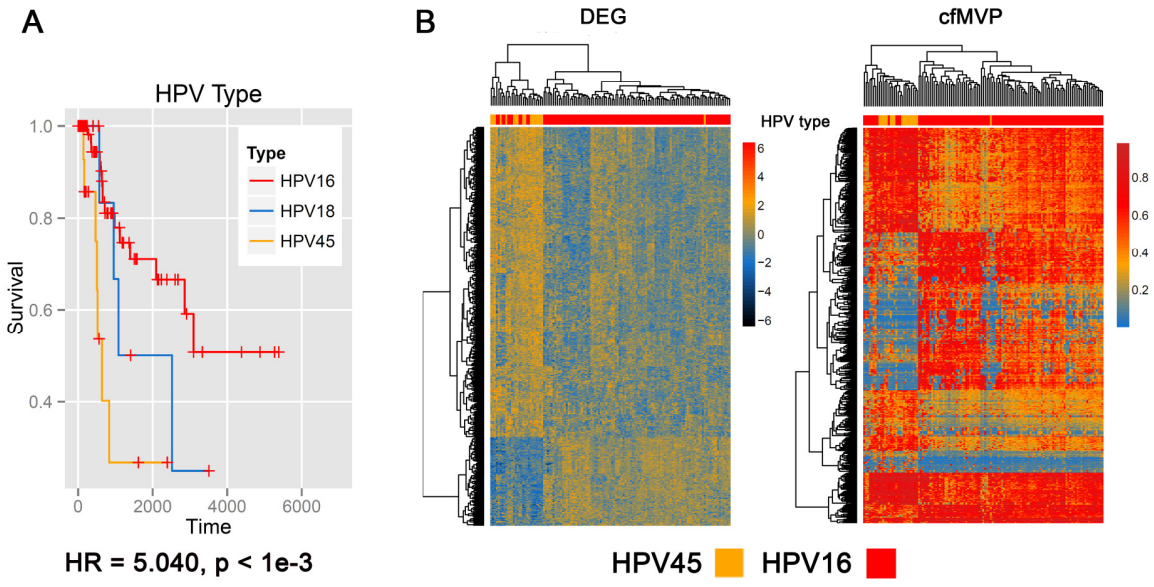


Figure 36: A) Survival differs by HPV type with HPV16 and HPV45 occupying opposite ends of the prognostic spectrum (Stage I and II tumours). X axis = time in days, Y axis = risk of death B) Heatmaps showing 646 DEG and 614 expression associated MVPs respectively in Stage I and Stage 2 squamous cell cancers (n = 127). Annotation ribbon depicts HPV type (HPV45 in orange, HPV16 in red).

In order to examine the interplay between taxonomic categories and clinical outcomes at higher taxonomic resolution, and to derive signatures for clinical stratification, I compared survival by type in early-stage tumours, and found that HPV16 driven tumours and HPV45 driven tumours occupied opposite ends of the clinical outcome spectrum, with HPV18+ tumours displaying intermediate survival trends **(Figure 36A)**.

Modelling transcriptional differences between HPV16 and HPV45+ early-stage tumours identified 646 genes differentially expressed between HPV45 and HPV16 positive tumours (FDR=0.01, FC > 2). Similarly, analysis of methylation data and subsequent integration with differentially expressed genes yielded a signature of 613 MVPs, of which 225 genes were also present in the transcriptional signature compatible with the canonical function of DNA methylation. Heatmapping and clustering highlighted general separation of HPV45 and HPV16 driven tumours using the cfMVP and whole-transcriptome DEG signatures **(Figure 36B)**.

Pathway analysis points to an inflammatory phenotype in HPV45 driven tumours.

IPA core analysis was then carried out to characterise pathway signatures while limiting the database to experimentally observed direct interactions. Some of the top canonical pathways included Granulocyte and Aggranulocyte Adhesion and Diapedesis, and amongst canonical pathways with activation z-scores and clear-cut activation states, one of the top hits for activation in HPV45+ tumours was TREM1 signalling, a pathway previously implicated as being essential for hepatocarcinogenesis (Wu, Li et al. 2012), associated with the secretion of multiple proinflammatory cytokines.

TREM1 signalling is known to be induced in tumour associated macrophages and the expression profiles for these HPV45+ tumours showed significant upregulation of the macrophage marker MCP1 and PTGS1 and 2, which are known to induce TREM1 activation (Yuan, Mehta et al. 2014). *In toto* the HPV45 signature, which contains multiple inflammatory cytokines including interleukins (IL11, IL18, IL1B, IL24, IL6, IL8) and other immune mediators (CCL2, CXCL2, CXCL3, CXCL5, TNF-alpha and TNFAIP6) points to the presence of an aggressive, inflammatory phenotype associated with HPV45.

The HPV45 transcriptional signature points to metastatic behaviour

Gene set enrichment using IPA's Diseases and Functions ontology identified cellular movement as the most activated pathway, with 85 out of 121 genes in the set expressed consistently with increased metastatic potential **(**gene set visualised in **Figure 37)**.

The genes overexpressed in this pathway with prometastatic associations included, amongst others, SNAI1, which is a master regulator of epithelial mesenchymal transition and is associated with worse prognosis and chemoresistance in metastatic cancers (Kaufhold and Bonavida 2014), fibronectin 1, which is known to trigger EMT-associated transcriptional cascades (Park and Schwarzbauer 2014), RHOF, which has been implicated in the formation of pseudofilopodial protrusions needed for invasion through basement membranes(Reymond, d'Agua et al. 2013), and VEGFC, a VEGF isoform implicated in the process of tumour-associated lymphangiogenesis (Skobe, Hawighorst et al. 2001) .

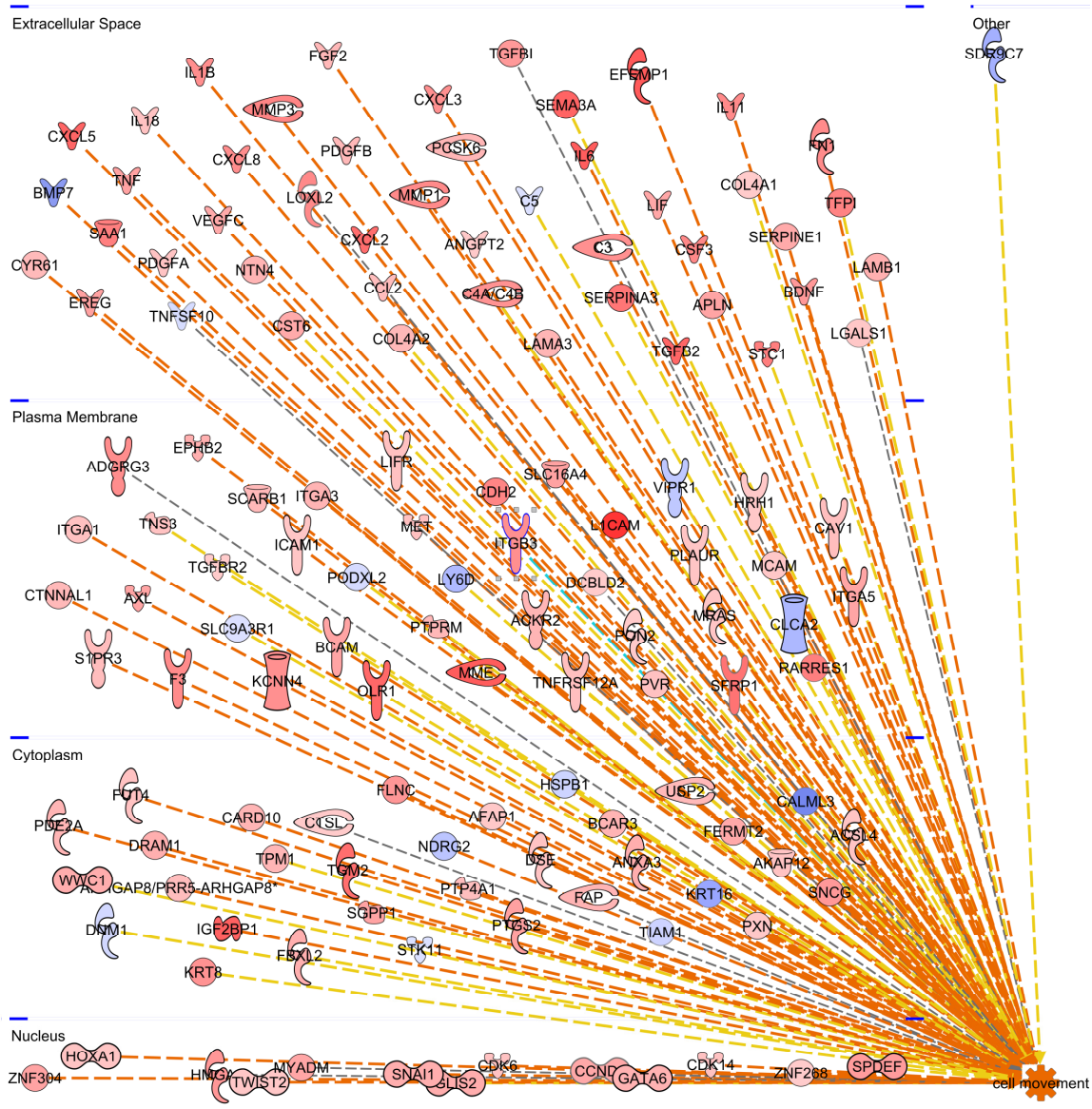


Figure 37: The "Cell Movement" gene set that is the top enriched gene set in HPV45+ tumours. Red = overexpressed in HPV45+ tumours, blue = underexpressed. Dark orange = expression change consistent with increased cell movement, light orange = expression change inconsistent with increased cell movement.

Upstream regulatory analysis identifies putative regulators of the HPV45 expression signature.

The top five candidate regulators inferred to be activated in HPV45+ tumours were SMARCA4, EZH2, RELA, FOXL2, and HIF1A. Interestingly, RELA is an NF-kB effector, and the co-ordinated activity of EZH2 and RELA in complex has been linked to the modulation of genes implicated in epithelial mesenchymal transition (Lee, Li et al. 2011), and some of the genes defined to be coregulated by EZH2 and RELA are also inferred to be coregulated by both by IPA (TNF,IL6, PTGS2).

Beta-catenin signalling was also inferred to be activated in HPV45+ tumours, and is consistent with the previously documented association of beta-catenin cellular localisation as a correlate of survival (Zhang, Liu et al. 2014). SMARCA4 (BRG1) interestingly has been implicated in the transcriptional activity of the HPV18 (He and Luo 2012) promoter before and it will be of interest to verify if the relative inferred activation of this gene may correspond with differential dependence by HPV type.

The functional context of HPV45 associated epigenomic changes.

IPA analysis of DMR-associated differential expression genes did not highlight many canonical pathways, upstream regulators or diseases and function associations of interest, leading us to focus on the much larger set of gene expression changes associated with MVPs. Upstream regulatory analyses again implicated SMARCA4 and CTNNB1 as being activated, and the aryl-hydrocarbon receptor to be inhibited.

Diseases and functions ontology analysis resulted in a list of associations including cell migration and proliferation amongst the top scoring hits for activated pathways/processes. These results lend an additional layer of support to the hypothesis that an aggressive, prometastatic signature characterises the behaviour of HPV45+ disease.

HPV45 associated molecular signatures are of value beyond HPV45.

It was apparent from the heatmaps presented in (**Figure 36B)** that certain HPV16+ samples clustered with HPV45+ samples, suggesting the HPV45 associated signature may be of relevance more broadly for outcomes in cervical cancers, including in stratifying the intermediate survival HPV18+ tumours by aggressiveness. The availability of both methylation and expression signatures, which have different distributions and scales demanded the development of a multiplatform clustering approach, which led to the use of a random forest proximity matrix to represent a joint Feature x Sample matrix consisting of probes and transcripts from both signatures. The joint clustering that resulted from this identified 7/18 HPV18+ tumours and a small subset of HPV16+ tumours (8/119), that clustered with HPV45 driven tumours **(Figure 38A)**. Comparing outcomes suggested these clusters retained prognostic performance overall (HR = 10.08, 95% CI = 4.07 - 24, q = 5.8e-7) when stratified by stage, and when restricted to early stage HPV16 and HPV18 tumours alone and stratified by stage (HR = 15.48, CI = 4.7 - 49.9, q = 6.9e-6), and when tested as part of a Cox regression model that was stratified by both type and stage (HR= 18.1, CI = 5 - 65.8, q = 1e-5). These signature derived pan-type clusters were therefore a good starting point for biomarker development to aid clinical stratification of patients.

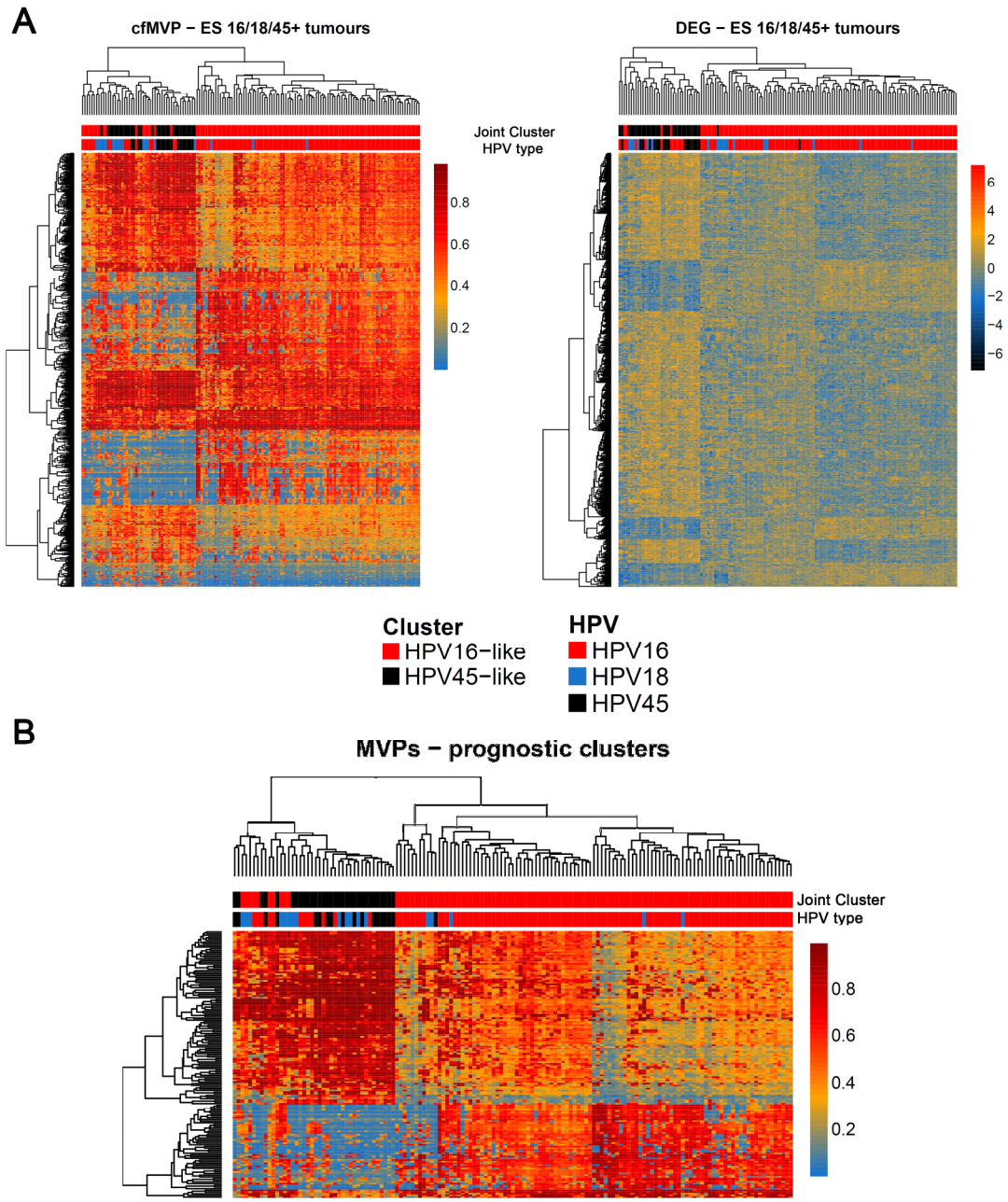


Figure 38: A) Heatmaps of HPV45-associated cfMVPs (left) and DEG (right) across all Early Stage I and II HPV16/18/45+ Cervical squamous cell cancers in the TCGA cohort. Cluster allocation is derived from proximity matrix of samples based on both data types. Multiple HPV16 and HPV18+ tumours display HPV45-like profiles. B) Heatmap of MVPs (delta-Beta 0.3, FDR < 0.01) between HPV45-like and HPV16-like clusters of Stage I and II cervical squamous cell cancers.

Machine learning yields a DNA-methylation classifier for aggressiveness associated clusters.

In order to develop a DNA methylation classifier for potential clinical application, I initially compared the HPV45-like and HPV16-like clusters of tumours and derived a signature of strongly differentially methylated probes (FDR < 0.01, beta value difference 0.3) **(Figure 38B)**. This was then used to train a set of classifiers (Random Forest, k-Nearest Neighbour, Support Vector Machine, Gradient Boosted Machine, Nearest Shrunken Centroid, GLMnet). Each model was evaluated using Kappa values derived from aggregates of out-of-fold predictions to estimate ability to recapitulate the original clusters. The Support Vector Machine was found to be the best model (Kappa = 0.96, PPV= 0.98 for HPV-16 like tumours, NPV = 1.00). The model was then applied to independent methylome datasets with accompanying clinical information from Norway (n = 87) to validate prognostic utility. The model stratified patients by survival after controlling for stage and histology with HPV45-like tumours exhibiting significantly worse outcomes (HR = 4.134, p = 0.048, 95% CI = 1.01 – 16.8) after controlling for stage (Kaplan-Meier Curve of the cohort is in **Figure 39**). Notably, amongst driver genes only *KRAS* and *PIK3CA* are mutated disproportionately (FDR < 5%) between the two clusters in TCGA, suggesting the bulk of prognostic differences are mediated by the epigenome and the transcriptome. It is notable that when Histology alone was evaluated as a mode of stratification while adjusting for stage, none of the histological subtypes attained significance at p < 0.05. At the time of writing, additional processing of samples from an Austrian cohort of Cervical cancers was underway to enable further validation of the aggressiveness signature in a more powerful cohort.

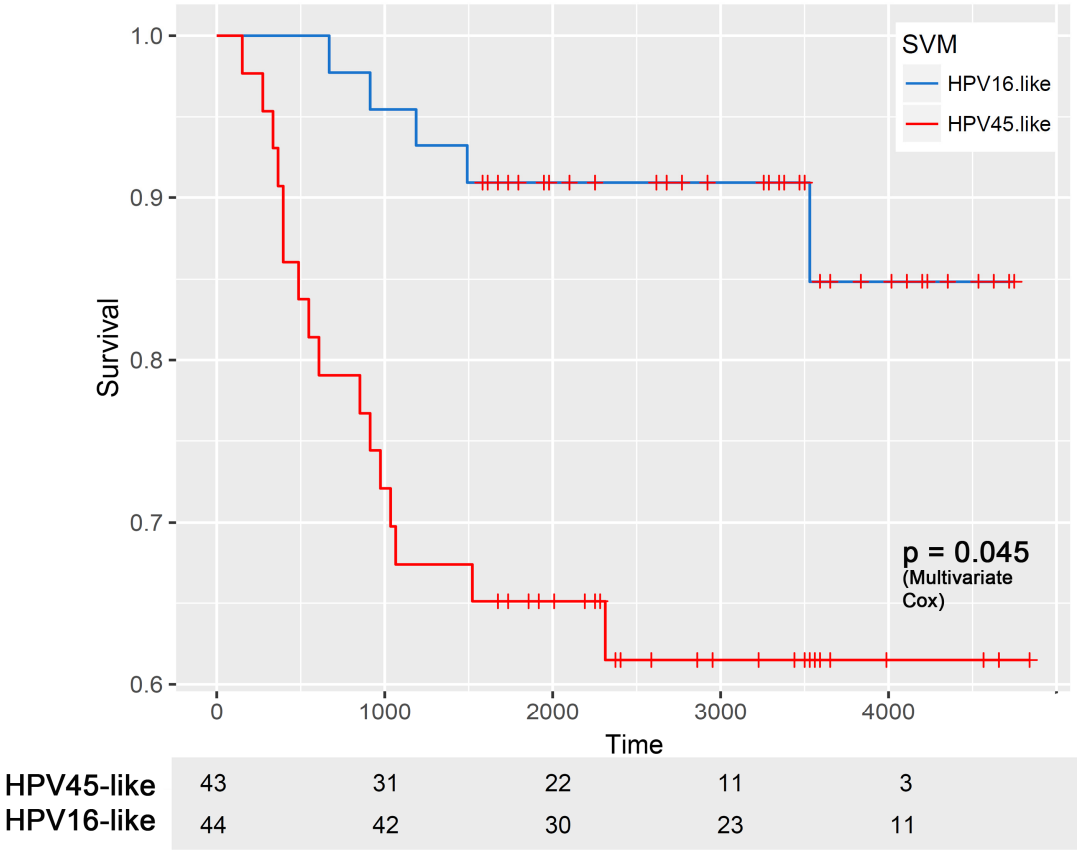


Figure 39: Kaplan-Meier Curves of predicted HPV45-like and HPV16-like tumours within the Norwegian Cohort of Cervical Cancers. X-axis = time in days. Appendix A5 plots the K-M curves by histology, serving to explain the marginal P-value observed here.

Aggressiveness Clusters Span Histology despite Cell-of-origin preferences.

Having established that HPV45 and HPV18+ tumours tended to display high-expression of Squamocolumnar-junction cell genes (Chapter 4) I tested for associations between Junction-signature cluster and Aggressiveness allocation, which revealed a strong tendency for HPV45-like tumours to also display high levels of squamocolumnar junction transcription (OR=4 for a Junction-High tumour to be HPV45-like, p = 7.64e-5, Fisher's Exact Test), suggesting that HPV45-like tumours may arise from a distinct cellular population to which HPV18 and HPV45 are restricted.

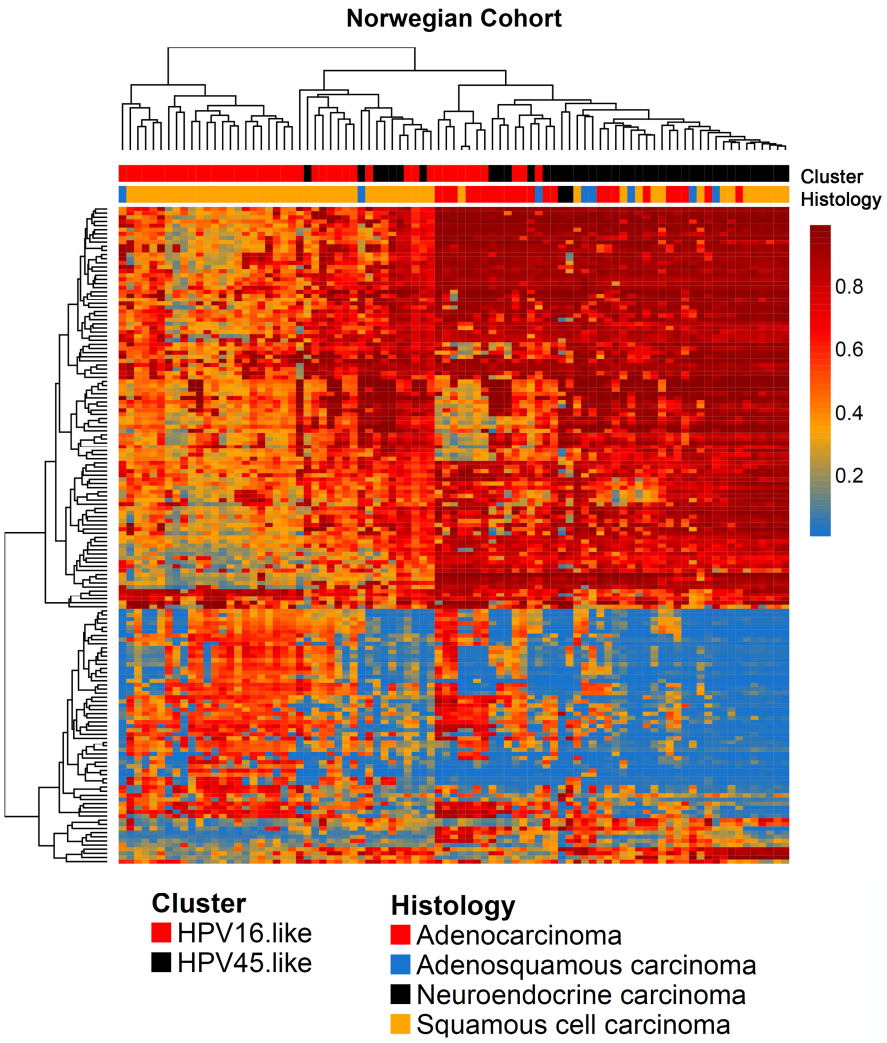


Figure 40: Heatmap of SVM-based Aggressiveness Class allocation in the Norwegian Cohort. Cluster = Allocated Class, Histology = histopathologic classification.

Visualisation of class allocation and histology in the Norwegian cohort indicated enrichment for adenocarcinomas in the HPV45-like cluster. Formally testing this hypothesis in the entirety of the TCGA CESC cohort indicated adenoid (adeno/adenosquamous) carcinomas were far likelier to be HPV45-like (36%) than squamous carcinomas (20%) (OR = 0.41, p = 0.01).

Chapter Conclusions

Cervical Cancers are unique amongst HPV+ tumours in terms of being driven by multiple common HR-HPV types, and previous research has hinted that different HPV-types may be associated with different molecular profiles and pathology. The work in this chapter firmly establishes the concept that different HPV-types, when grouped on the basis of taxonomic categories such as clade/species and HPV type, are associated with large-scale transcriptional and epigenetic differences, either as the proximate cause or as a consequence of cellular background and wiring. HPV45+ tumours were particularly aggressive in the TCGA CESC cohort and displayed molecular profiles that had prognostic value across the cohort, culminating in the development of a classification scheme that stratifies patients by prognosis. Additional validation confirmed the existence of these subtypes in an independent cohort and again hinted at the markedly aggressive nature of HPV45-like tumours, in a single-center cohort with well-annotated clinical data, which works-around the multi-centre nature of TCGA clinical data and the lack of cause-specific mortality information in the latter.

Statistical analysis suggests that HPV45-like tumours are often enriched for a Squamocolumnar Junction gene expression profile and the molecular profiles that underpin them cut across lines of histology. This offers a new approach to classifying and managing cervical cancers that goes beyond and complements constructs like staging and histology that currently define clinical management. A critical open question at this point, meriting future investigation, is whether different HPV types are sufficient causes for the induction of these molecular profiles.

The Immune Landscape of HPV-driven tumours.

The analyses carried out in this chapter are summarised in a flowchart in Appendix A3

Background

Findings presented in Chapter 3 demonstrated that the immune system could be a potential mediator of prognosis amongst subsets of HPV-driven tumours that are otherwise molecularly similar. In Chapter 6 I demonstrated the existence of a distinct set of molecular signatures associated with HPV45 with transcriptional patterns suggestive of pro-inflammatory signalling and TREM1 activation, and then demonstrated that these patterns were also associated with subsets of HPV18+ and HPV16+ cancers.

The viral origin of these tumours suggests that as they develop, they must evolve to evade recognition by the immune system. Multiple lines of evidence support this assertion – the markedly increased risk of HPV-associated malignancies in HIV/AIDS patients with immune system dysfunction(Palefsky 2009), amplifications in antiviral genes that mediate episome clearance (Mine, Shulzhenko et al. 2013), the association between anti-HPV16 E6 antibodies and risk of full blown malignancy in candidate patients for HPV+ OPSCC and to a lesser extent in genital cancers (Kreimer, Brennan et al. 2015), the association of germline variation in MHC and HLA genes with risk for cervical cancer (Chen, Cui et al. 2015) .

In addition to this, a substantial body of work has emerged around how the evolution of tumours is shaped by the immune system through the process of immunoediting (Mittal, Gubin et al. 2014), where the immune system acts to exert selective pressures on mutants and novel antigenic proteins to select for cells that either have dampened antigen presentation or eliminate highly antigenic variants to produce cellular populations capable of escape.

Evidence has steadily accumulated to suggest that measures of immune activity can be associated with prognosis and that signaling through specific pathways in cancer cells can result in alterations in external immune function, and that the presence of distinct antigenic and mutational signatures can be indicative not only of prognosis but of response to immunotherapies that enable the immune system to counter tumours that have, through prior immunoediting, evolved to convert an anti-tumour immune microenvironment to a pro-tumour immune microenvironment (Snyder, Makarov et al. 2014, Rizvi, Hellmann et al. 2015, Hugo, Zaretsky et al. 2016).

Multiple methods have been recently developed to permit the in-silico deconvolution of complex cellular mixtures and to estimate tumour purity, be they reference-free methods (Houseman, Molitor et al. 2014) or reference-based methods (Koestler, Christensen et al. 2013, Jaffe and Irizarry 2014). Methylation data particularly have been demonstrated to be amenable to deconvolution of complex tissue mixtures, as demonstrated in the deconvolution of blood.

In this chapter, I document the development of a cellular deconvolution approach on bulk tumours using methylation data and combine it with the results of the previous chapters and molecular classification to delineate the role of the immune system in the context of HPV-driven cancers, including in light of the aggressiveness based classification previously derived.

Methylation modelling yields an accurate cellular deconvolution approach based on Support Vector Regression.

A variety of methods have been developed to deconvolute complex cellular mixtures into component cellular fractions using methylation data, mostly in the context of blood. Support Vector Regression has previously been used with expression data and bulk tumour transcriptomes to break down the haematopoietic fraction of cells into component populations (Newman, Liu et al. 2015). In order to deconvolute bulk methylomes into tumour, stroma and immune cell subtype-associated components, I assembled and curated a collection of methylomes consisting of HNSC and CESC cell lines to serve as the tumour reference, fibroblast samples to serve as stromal representations, and eight distinct cell types and performed feature selection to identify a total of 510 probes to discriminate between these cell types using support vector regression (referred to from this point onwards as MethylCIBERSORT) **(Figure 41A)**.

The accuracy of the estimates was gauged by comparison to previously published ABSOLUTE-estimates of tumour-purity available for 466 HNSC. ABSOLUTE (Carter, Cibulskis et al. 2012) uses mutation/copy-number data to jointly optimize the best model for purity, ploidy and observed variant allele frequencies and has been demonstrated to be highly accurate in previously published work by comparisons with flow-cytometry analysis. A high magnitude of correlation was revealed between MethylCIBERSORT estimates and ABSOLUTE purity estimates ($R = 0.79$, $p < 2.2e\text{-}16$, Spearman's Rank Correlation), confirming the utility of the method in deconvoluting fractions within bulk tumour samples **(Figure 41B)**.

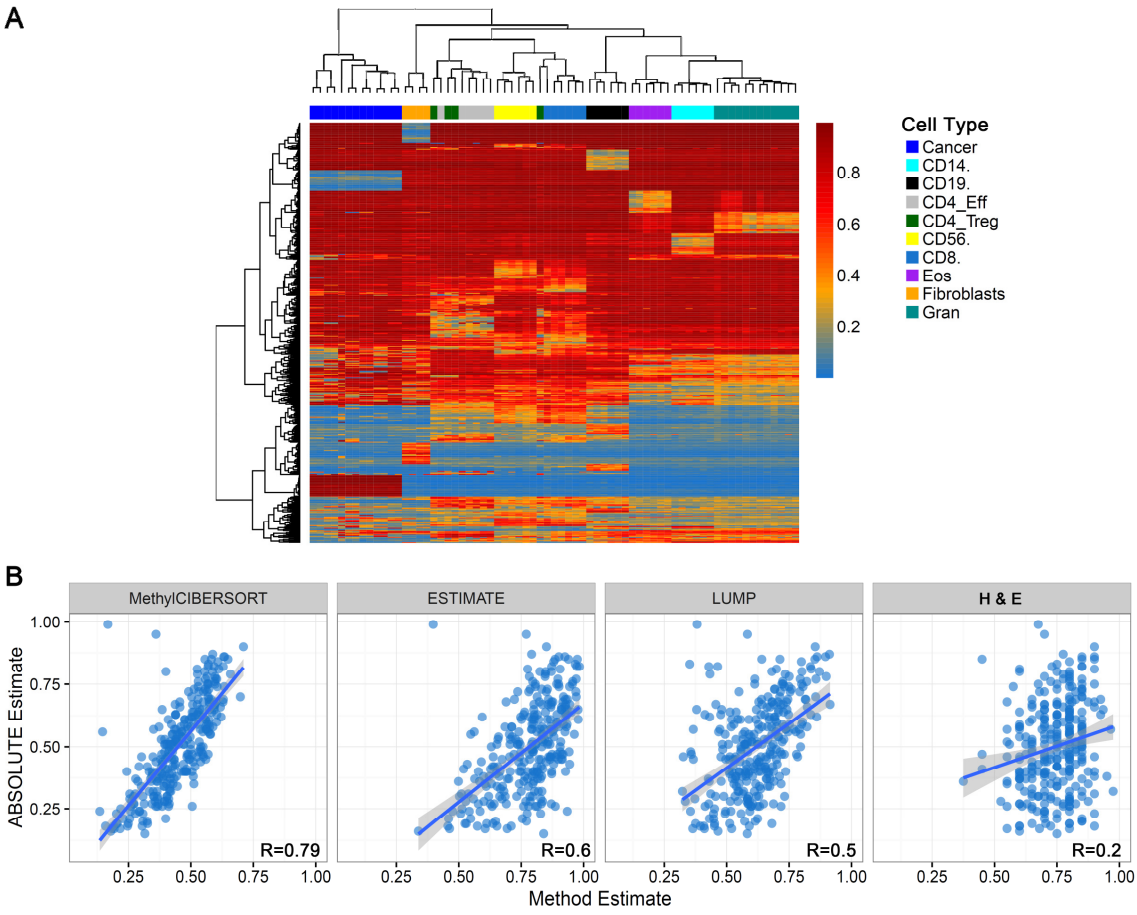


Figure 41: A) Heatmap showing cell-type specific markers used for deconvolution, rows are features and columns are samples, B) Comparisons for estimated cancer cell content between ABSOLUTE (Reference Gold Standard) and other methods. Y-axis = ABSOLUTE cancer cell content, X- asix, method specific estimates. R= Spearman's Rho. MethylCIBERSORT outperforms previously published methods. Data for histopathological estimates, ESTIMATE estimates and LUMP for HNSC samples retrieved from (Aran, Sirota et al. 2016).

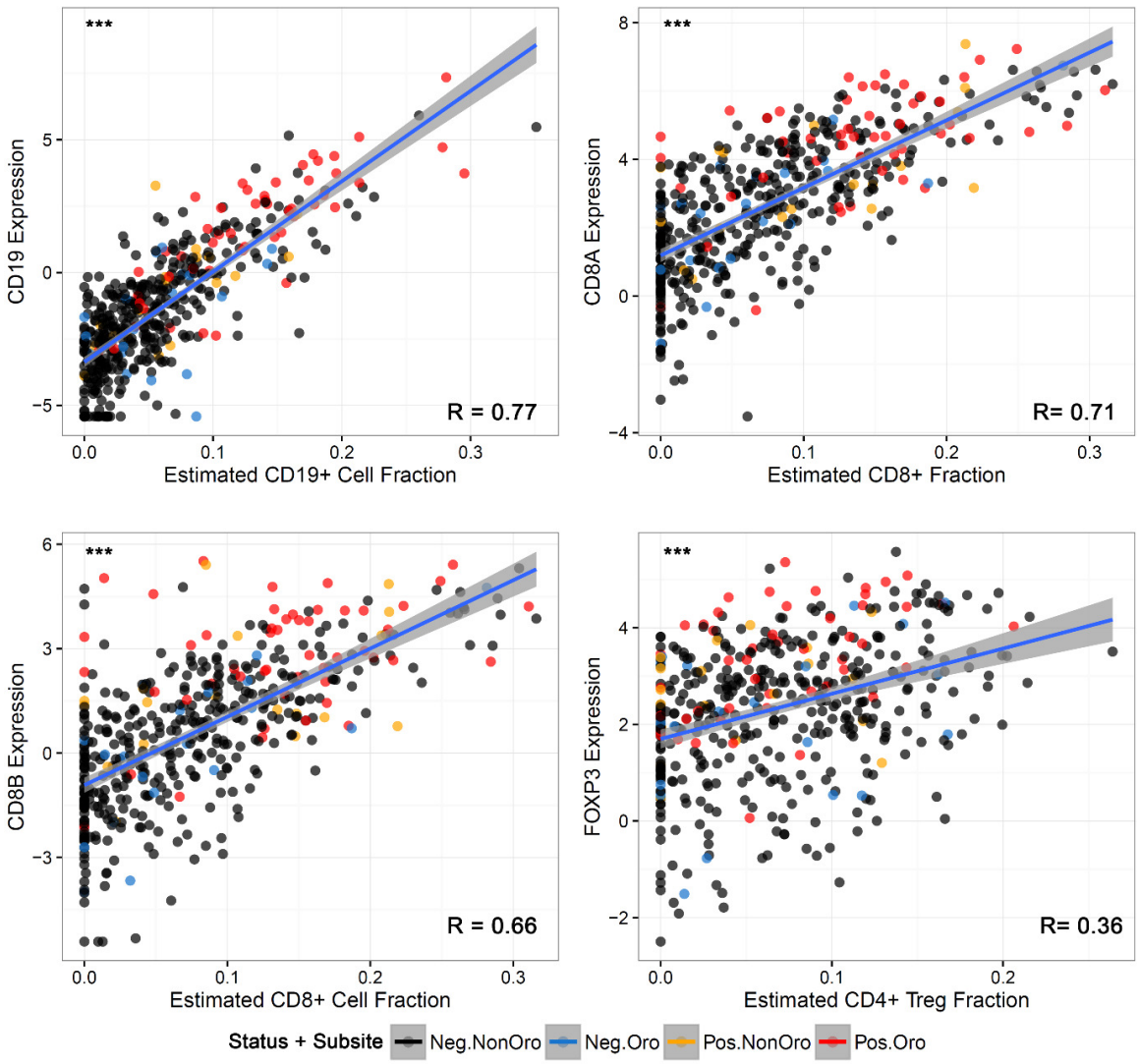


Figure 42: MethylCIBERSORT estimates are significantly correlated with expression of marker genes. Y Axis = Transcript counts (log2 cpm), X axis = MethylCIBERSORT estimate. * = BHFDR < 0.01, R = Spearman's Rho. Estimates from 466 HNSC coloured by Anatomic subsite and HPV status to permit comparison with transcript-wise distributions visualised in Chapter 1.**

Notably, MethylCIBERSORT outperforms previously published methods of estimating tumour purity such as LUMP (Aran, Sirota et al. 2015), which also uses methylation data (R=0.51 with ABSOLUTE, p < 2.2e-16) and ESTIMATE (Yoshihara, Shahmoradgoli et al. 2013)(R= 0.54, p < 2.2e-16) **(Figure 43B)**. Additional validation was carried out by estimating correlations between transcription of key marker genes (CD19 for B Cells, FOXP3 for Tregs, CD8A/CD8B for CD8+ lymphocytes) and the estimated immune cell fractions. In every case, significant correlations (Range – 0.36 to 0.77) were observed between transcript levels and estimated infiltrate. Taken together, these findings establish MethylCIBERSORT as a useful approach for probing immune cell content in admixed tumours and facilitate deep deconvolution of tumour cell components, as opposed to simple division into tumour, immune and stromal components.

Variability in Immune Cell Content is associated with Cervical Cancer Aggressiveness.

In Chapter 7, I identified patterns that divided cervical cancers into two subgroups with distinct prognoses. The SVM classifier developed using early stage tumours was then applied to the entirety of the 844 methylomes assembled to serve as the discovery set in Chapter 5. Cell fraction estimates were compared for all HPV+ CESC subsequently and significantly lower fractions of Natural Killer cells (CD56+, Median FC = 0.59), monocytes (FC = 0.72) and Granulocytes (Median FC= 0.59), and elevated levels of CD8+ lymphocytes (Median FC = 1.54) (all at FDR < 0.02, Wilcoxon's Rank Sum Test) were identified in HPV16-like tumours relative to HPV45-like tumours **(Figure 43)**.

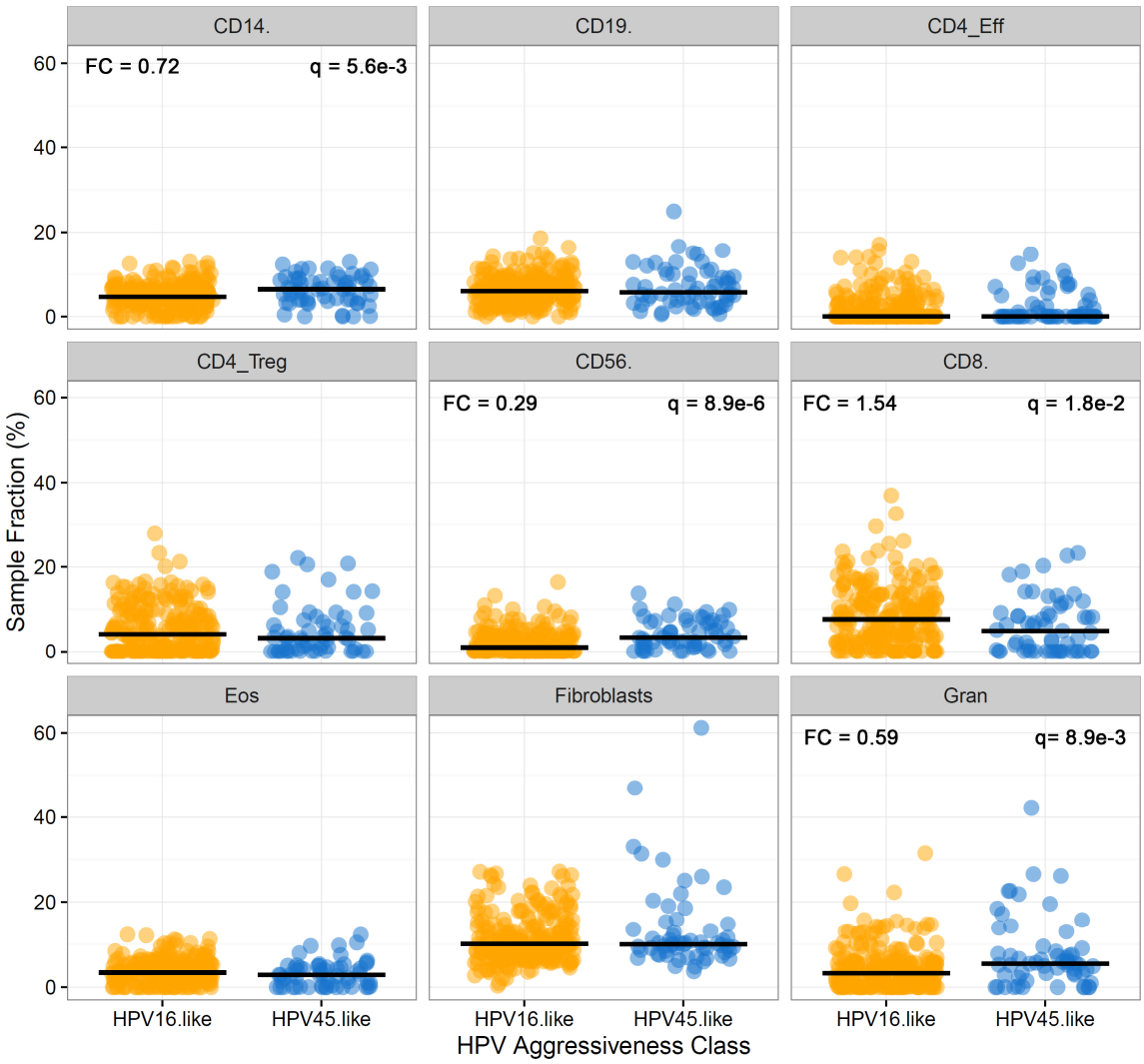


Figure 43: Estimated Immune Cell Fractions for different Immune Cell Types vary with Cervical Cancer Aggressiveness Cluster (All TCGA CESCs). *= FDR < 0.05, * = FDR < 0.001. FDR computed using BH-correction from Wilcoxon's Rank Sum Test P-values. Horizontal bars represent groupwise medians. X axis = Aggressiveness Cluster. Y axis = Estimated fraction for cell type in facet. Fold Changes defined as fraction in HPV16 relative to HPV45.**

The prognostic utility of high neutrophil : lymphocyte ratios has been previously documented in large cohorts of cervical cancers (Lee, Choi et al. 2012, Mizunuma, Yokoyama et al. 2015) and Neutrophils are the most abundant subset of Granulocytes. Consequently, in this context, infiltrating cell estimates implicate putative mediators of link between molecularly-defined profiles and outcomes.

NK-cells have previously been implicated as a pro-survival factor in multiple cancers, raising questions of why high levels of this cellular subset are a feature of these aggressive tumours. In HPV45+ tumours that served as the reference point for the definition of the signatures behind the original class allocation, it is notable that in TGFB1 (> 2FC) and TGFB2 (> 7FC) expression are highly upregulated, and TGF-Beta is known to induce suppression of NK-cell activity, potentially resolving this conflict (Chang, Li et al. 2016).

Integrative Analysis of Immune Cell Signatures in HPV+ tumours.

A pairwise correlation analysis was then carried out to examine if there were statistical associations between different infiltrating cell types in the 356 HPV+ tumours (CESC and HNSC) that were methylation profiled by TCGA. Significant associations were noticed for 19/36 pairs of infiltrating cell-types (Spearman's Rank Correlation, FDR < 0.01) **(Figure 44)**. This suggested that distinct subsets of infiltrating cells could potentially be associated with diverse biology in these tumours. In order to fully integrate cellular abundance estimates with immune function, I carried out joint-clustering of these estimates with lymphocyte effector markers using a combination of PAM clustering and a Random Forest dissimilarity matrix. This identified two major clusters, which displayed significant differences in the distributions of abundance estimates for 8/9 cell types (FDR < 0.05, Wilcoxon's Rank Sum Test). One group (C1, aka Immunoreactive) was characterised by high levels of CD4+ Tregs, CD8+ Lymphocytes, CD19+ lymphocytes and NK-cells whereas the other (C2, Immunodepleted) was characterised by high levels of fibroblasts, Granulocytes and CD4+ non-Treg cells **(Figure 45)**.

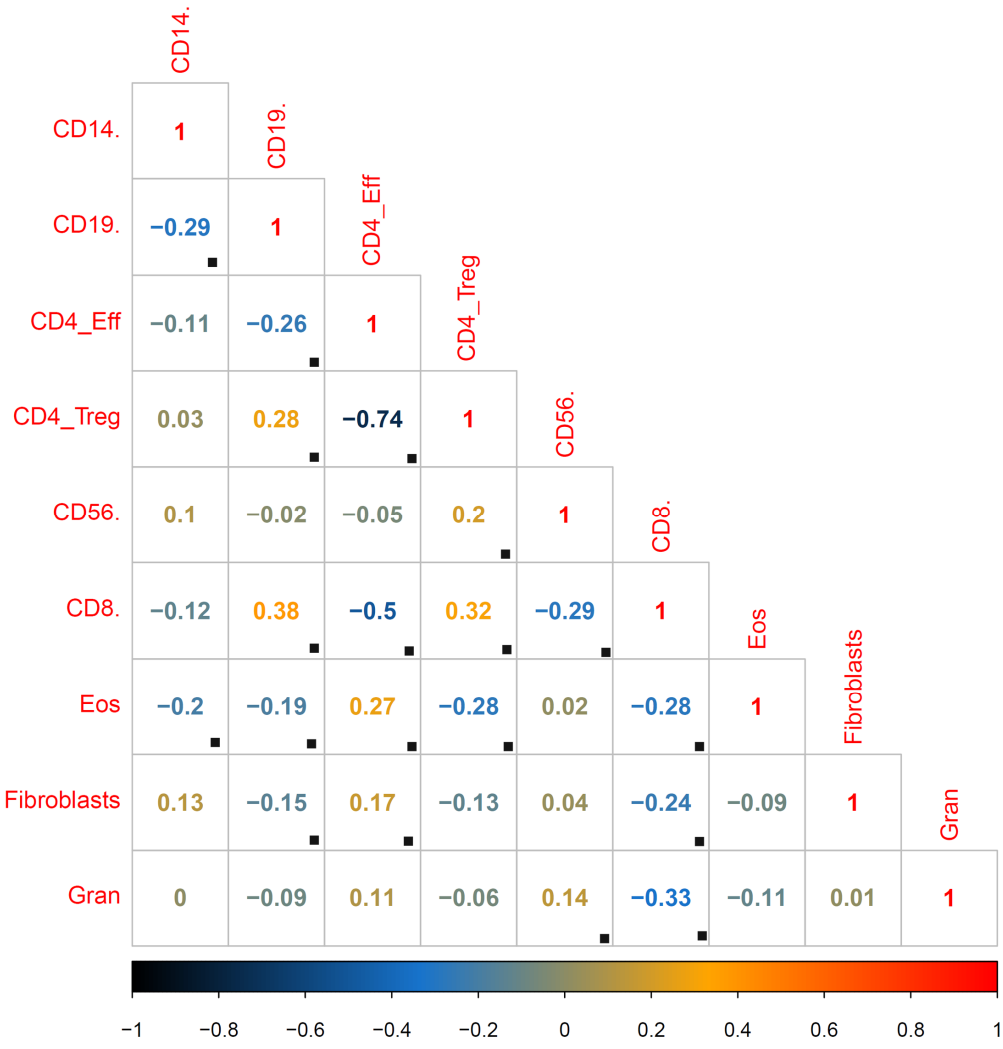


Figure 44: Pairwise correlation matrix of infiltrating cell type abundances estimated using MethylCIBERSORT in 356 HPV+ HNSC and CESC highlights multiple significant associations. Numbers in cells represent Spearman's Rho. Back boxes at bottom corners indicate statistically significant correlations at FDR < 0.01

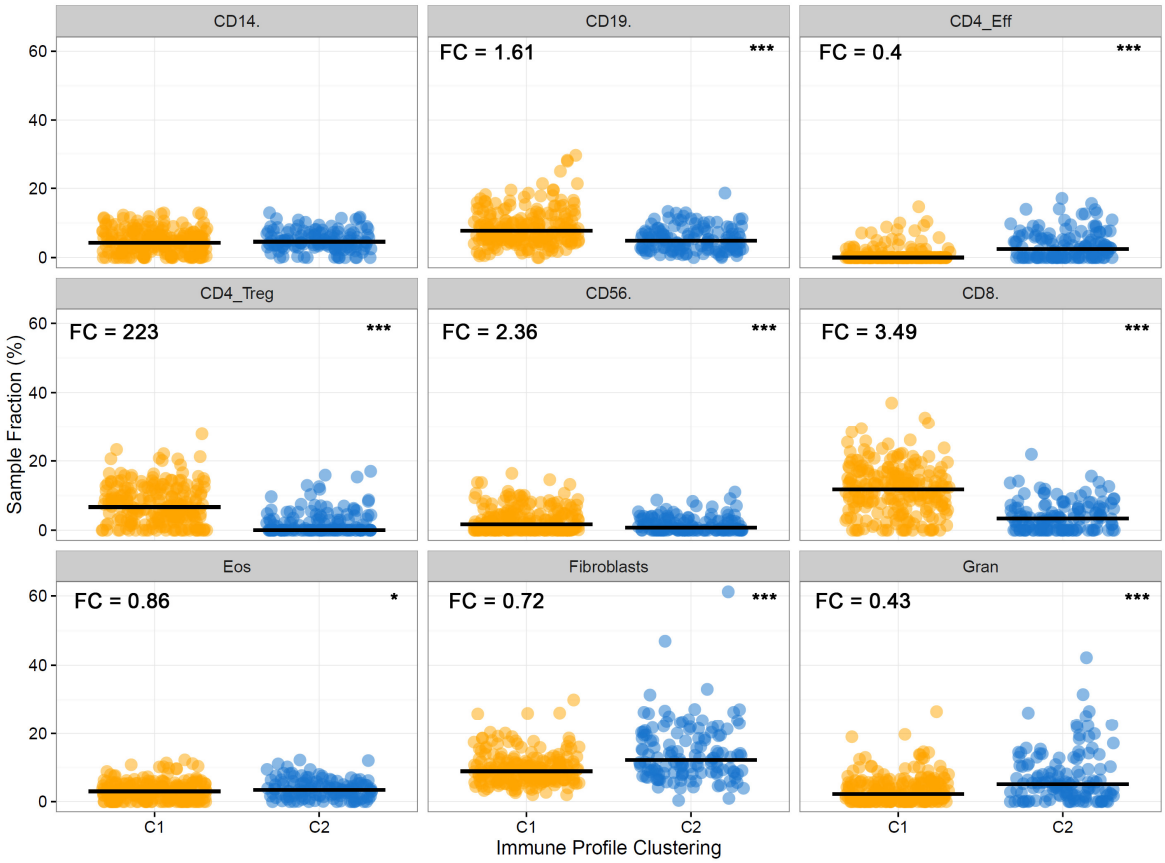


Figure 45: Distributions of Immune cells vary markedly by Immune Cluster. Y axis = estimated fraction per sample. X axis = Cluster. Crossbars represent group medians and colours represent overlaps with HPV Aggressiveness Classes. Fold changes calculated as medians of Cluster1/Cluster2. * = FDR < 0.05, * = FDR < 0.001**

In order to understand the biological significance of these two distinct patterns of infiltration and evaluate if they are associated with distinct patterns of molecular alterations, I performed comprehensive analyses of differences in gene expression, protein abundance / activity, and DNA methylation patterns with a functional impact on shaping gene expression between the two immune clusters. Limma-trend analysis of expression data identified 1176 DEGs (FDR < 0.01, > 2FC) between the two clusters **(Figure 46)**.

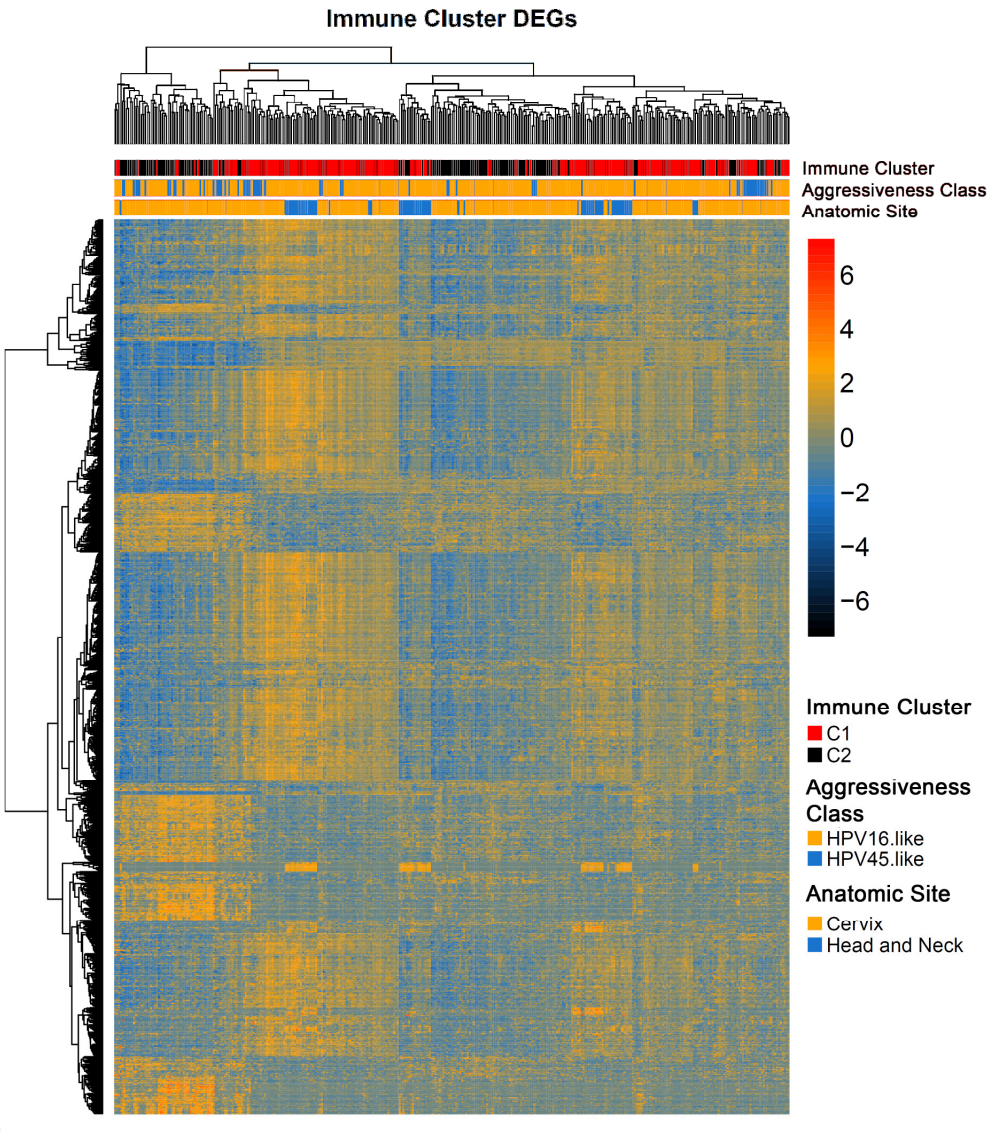


Figure 46: Heatmap of 1176 genes differentially expressed between the two Immune Clusters (FDR < 0.01, 2FC). Rows represent genes, intensities represent expression Z-scores and columns represent samples.

Pathway analyses were highly reflective of differential immune activity, with the top 5 Canonical Pathways including Crosstalk between Dendritic Cells and NK Cells, T-Cell Differentiation, Altered T-Cell and B-Cell signaling in Rheumatoid Arthritis, Natural Killer Cell Signalling, ICOS-ICOSL signaling in Helper T-cells and Allograft Rejection Signalling (All at BH FDR < 0.01).

Diseases and functions ontology analysis again identified decreased leukocyte and lymphocyte migration and activation in Immunodepleted samples and identified elevated activity linked of the gene set titled “Infection of Virus” in this cluster. Reassuringly, *CD8A*, the Treg-markers *TIGIT* and *FOXP3*, the NK-cell marker *NKG7* and *CD19* were all significantly underexpressed in the Immunodepleted cluster as expected from analysis of cluster-wise cellular estimates.

Upstream regulatory analysis identified differential regulation of multiple key Interferon-response transcription factors. IRF1, IRF3, IRF5 and IRF7 were inhibited in this cluster while IRF4 was activated. HPV16 E7 has been shown to directly bind to IRF1 and recruit HDAC to IRF downstream genes to inactivate transcription (Park, Kim et al. 2000), and HPV16 E6 has been shown to interact and block the activity of IRF3 (Ronco, Karpova et al. 1998), making these inferences consistent with HPV biology.

In addition to interferon response genes, several pattern recognition receptors – *TLR7*, *TLR8*, and *TLR10* are strongly upregulated in Immunoreactive tumours. *TLR7* and *TLR8* specifically have been shown to be associated with Langerhans Cells (Fahey, Raff et al. 2009), suggesting that some of the aforementioned expression changes may indeed be due to changes in the microenvironment. Other upstream regulators systematically dysregulated include the products of *NFkB2* and *RELA*, *FOS*, *EHF*, *SMARCA4*, *EZH2* and *SPI1* (activated in Immunoreactive tumours) and those of *ESR1*, *ID2* and *MSC* (activated in Immunodepleted tumours).

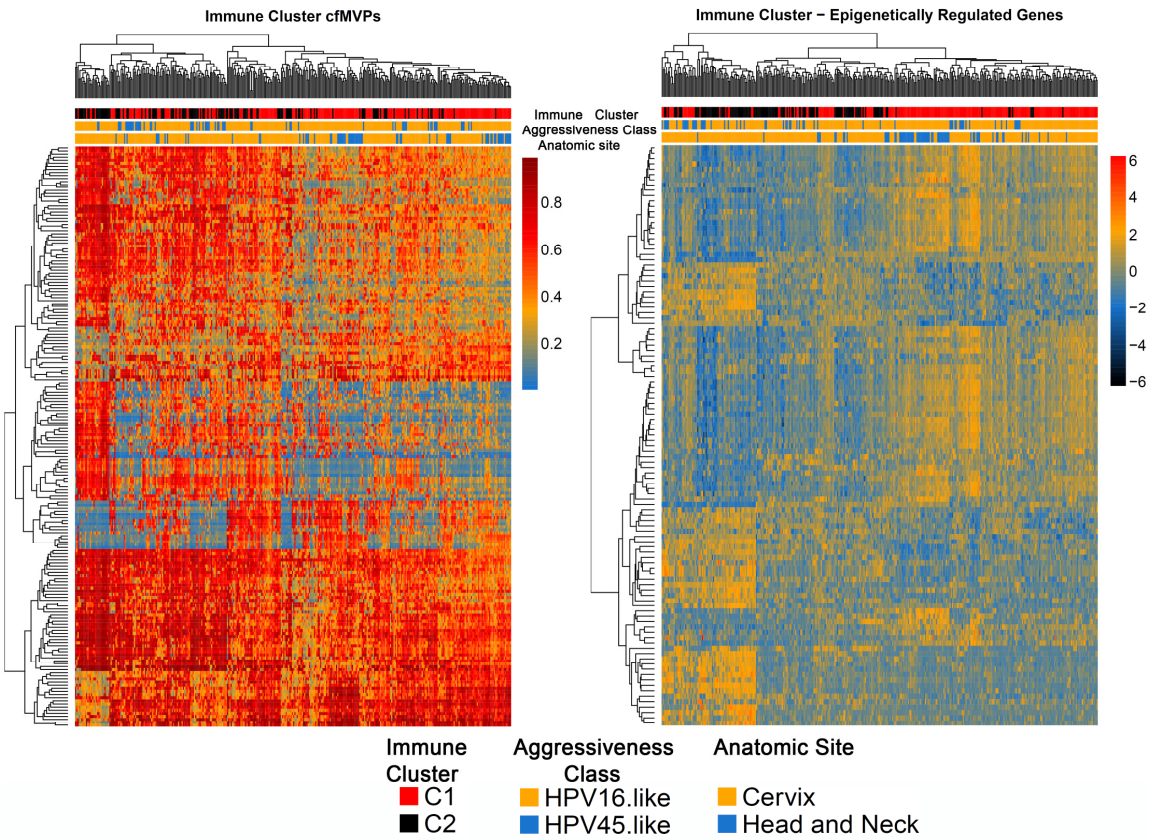


Figure 47: Heatmaps of cfMVP Beta-values (left) and expression of cfMVP-associated genes (right). Rows represent features and columns represent samples. Annotation ribbons represent Immune Cluster, Aggressiveness Class and Anatomic site respectively. See text for definition of cfMVPs and DEGs.

DNA methylation analysis was also carried out to estimate if underlying DNA methylation changes could be driving disparities between the two clusters and 190 cfMVPs (median deltaBeta = 0.15, FDR < 0.01), mapping to 109 DEGs, were identified to be differentially methylated between the two Immune clusters **(Figure 47)**. Upstream regulatory analysis inferred inhibition of IRF7 in Immunodepleted tumours, with *CCL5, IFIT1, STAT1* and *TRAF1* all epigenetically silenced.

The suppression of *STAT1* expression through HPV E6 and E7 has been demonstrated to be integral for genome amplification and plasmid maintenance of HPV16 in keratinocyte models (Hong, Mehta et al. 2011) and further reinforces the theme of HPV biology shaping molecular profiles that has been established in this thesis. STAT1 in mice has been shown to play a potent role in increasing antigen presentation when cancer-cell intrinsic and has been associated with the full activation of infiltrating adaptive immune cells when extrinsic (Meissl, Macho-Maschler et al. 2015), and may be an example of epigenetic events playing a role in immune evasion. These epigenetic events, along with those regulating the expression of checkpoints like *HAVCR2* (Anderson 2014) *,* cytosolic DNA sensors like *AIM2* (Man, Karki et al. 2016), and components of MHC class II molecules such as *HLA-DPA* and *HLA-DPB* may be targets for potentiating immunotherapeutic interventions through the pharmacological modulation of DNA methylation.

Analysis of RPPA (Reverse Phase Protein Array) data identified differential antibody binding for 11 proteins between the two immune clusters (FDR < 0.01). IGFBP2, Tafazzin, Fibronectin and AMPK were upregulated in the Immunodepleted group and Lck (lymphocyte specific, (Moogk, Zhong et al. 2016)),Bim, PREX1,STAT5-alpha and cleaved Caspase 7 were upregulated in the Immunoenriched group. The limited coverage of the RPPA platform (192 antibodies) and the small number of samples (n = 162, all CESC) represented render functional inferences challenging. A key limitation of these analyses is it is difficult to disentangle the transcriptional, epigenomic and proteomic contributions of tumour cells from those of immune cells. Molecular profiling of dissociated tumour and immune cells is an experimental step that will facilitate this.

Survival analyses of Immune Clusters suggests they comprise prognostically equivalent states of the Immune microenvironment.

Finally, survival analyses were carried out to estimate the contributions of immune clusters to survival using Cox models containing Immune cluster, age, and stage across the cohort of HPV+ tumours for which clinical data were available (n = 337), separately for the Head and Neck and the Cervix.

The immune signature did not display significant associations in either setting (HR = 1.37597, p=0.22 for CESC, HR = 2.3, p = 0.34). This finding suggests that the two distinct immune environments do not make significantly large contributions to survival differences once other clinical factors have been taken into consideration in the setting of traditional therapeutic modalities.

This was surprising given the prognostic value of the Immune-panel cluster derived in Chapter 3, which strongly suggested that HPV-driven cancers could be subject to immune activity in prognostically-relevant ways. This led me to evaluate alternate Cox models regressing measures of cellular abundance along with clinical covariates using two distinct stratification approaches.

One involved estimation of coefficients per percent increase in infiltrating cell abundance while including all available infiltrating cell types with age and stage as clinical covariates, while the other involved stratifying into quartiles and evaluating the prognostic difference between the top and bottom quartiles with the aforementioned covariates (but only evaluating one cell type at a time).

Cell Type	CESC Key – HR (P-value)	HNSC Key – HR (P-value)
CD14	1.11 (p = 0.024)	0.65 (p = 0.10)
CD19	0.90 (p = 0.024)	1.001 (p = 0.98)
CD8	1.05 (p = 0.043)	2.04 (p = 0.034)
Granulocytes	1.04 (p = 0.1)	3.88 (p = 0.01)
CD4+ Treg	0.95 (p = 0.28)	1.75 (p = 0.032)
CD56 (NK Cells)	0.99 (p = 0.99)	1.14 (p = 0.47)
Eosinophils	1.09 (p = 0.16)	4.14 (p = 0.029)
Fibroblasts	1.04 (p = 0.069)	0.96 (p = 0.82)
CD4+ Helper Cells	1.04 (p = 0.04)	0.19 (p = 0.03)

Table 7: Table of Cox regression coefficients and P values from additive model with all infiltrating cell types with age and stage as covariates. Significant associations in blue text. Hazard Ratios are per percent increase in estimated fraction of the corresponding cell types.

Across the two anatomical subsites, different cell types were associated with different Hazard ratios after controlling for other cell types and some cell types were significant specifically in one anatomic site but not the other **(Table 7)**. It is however difficult to make direct comparisons across anatomic sites. In models that evaluated differences between top and bottom quartiles, there were no significant associations found within HNSC, while CD4+ Tregs (HR = 0.45, p = 0.03) and CD19 (HR = 0.37 , p = 0.01) were found to be significant predictors in CESC. A comparative evaluation of the different survival modelling strategies is beyond the scope of this thesis.

Chapter Conclusions

In this chapter, I developed and characterized a method for estimating the cellular abundances of different infiltrating cell types in tumours using methylation data and demonstrated that accurate estimates and inferences can be drawn using the MethylCIBERSORT approach. As more methylation profiles are generated from cancer cell lines and immune subtypes with different tissues of origin, it will be possible to perform effective cellular deconvolution at higher resolution across a broad spectrum of cancer types. This should facilitate investigation of the immunological repertoire of cancers on a pan-tissue basis.

Differences in infiltrating cell fractions are associated with the Aggressiveness Classes/Clusters of Cervical Cancer defined in previous chapters on the basis of HPV-type associated molecular variation and may explain some of the prognostic differences between these classes.

An integrative clustering approach identified two distinct patterns of immune infiltration and activity with epigenetic, transcriptomic and proteomic correlates that may both serve as alternate mechanisms of immune evasion that prevent tumour destruction. Some of these alterations, especially epigenetic ones, could be amenable to therapeutic targeting to potentiate anti-tumour immune responses.

Finally, novel associations were uncovered between abundances of different cell types in tumours defined by anatomic site, serving as a starting point for generating hypotheses into immune dysregulation in these cancers. Unanswered questions at this point include whether episomal and integrant HPV-driven tumours demonstrate alternate mechanisms for immune evasion, and the role of mutational repertoire and other genomic alterations in determining immune cluster membership and antitumour immunity.

Discussion

Pan-tissue transcriptional similarities unify HPV driven cancers.

At the outset of this thesis, the evidence supporting the existence of HPV-specific, tissue-independent molecular profiles was sparse and fragmented. Integrating multiple collections of HPV transcriptional profiles led to the development of the most comprehensive transcriptional signature for HPV-driven carcinogenesis to date, which appeared to be generated by a combination of HPV oncoprotein expression and additional cellular changes in this context.

While some of the evidence showing this in transformed keratinocytes and transformed Mesenchymal Stem Cells was limited by the lack of HPV- tumour controls, the original process for deriving the metasignature and the validation of the metasignature using machine learning approaches in the TCGA cohort (n=566 samples, 68 HPV+) established the specificity of this transcriptional signature to tumours actively expressing HPV oncoproteins, bypassing the caveats associated with analyses limited to HPV DNA+ tumours that may not be actively transcribing HPV.

Many of these changes, upon functional annotation, reflected classically established facts in the biology of HPV and point towards HPV as the key player in shaping the evolution of these patterns.

Examination of these expression patterns across multiple cancer types revealed the magnitude and the involvement of multiple HPV-associated transcriptional changes of a large subset was unique to HPV-driven tumours.

Successful identification of a driver role for HPV in non-Oropharyngeal HNSCs uncovers a role for immune response in mediating outcomes.

There have been outstanding questions in the field regarding the contribution of HPV to the establishment of tumours in HNSCs that lie outside the Oropharynx, where HPV DNA is found at low frequencies compared to OPSCCs which display high aetiological burdens for HPV, as well as to prognosis, which is excellent in HPV+ OPSCC but is vastly more ambiguous outside the Oropharynx.

Leveraging on the ability to identify active HPV-transcription, and applying classifiers trained on pan-tissue signatures of HPV-induced transformation derived earlier in the thesis leads to the conclusion HPV plays a driver role in non-OPSCCs that display active HPV transcription but show outcomes more comparable to HPV- tumours because of differences in immune infiltration, which corresponds to differential transcription of CD8 effector molecules and is marked by a distinct expression pattern of immune checkpoint transcripts which overall split HPV+ HNSCs into groups with different outcomes .

These findings suggest a pathway towards patient stratification, where molecular markers are used to classify tumours by HPV-status initially, followed by an evaluation of TIL infiltration levels or immune checkpoint expression, in order to select patients for treatment with immunotherapy, which I hypothesize will demonstrate efficacy especially in TIL-High and Checkpoint-High tumours.

APOBEC-mediated mutagenesis links transcriptional and genomic evolution of HPV+ tumours.

It has long been known that additional genomic events are required to convert cells that express E6 and E7 constitutively to fully malignant cells, but it was unclear if HPV played an active role in driving these evolutionary processes or whether the acquisition of additional hits was a statistically independent process. While E7 induced chromosomal instability and the abrogation of p53 permitting evasion of cell-cycle arrest have been previously documented, it has been unclear what role HPV may play in shaping the evolution of single nucleotide variants in these tumours.

The derivation of the pan-tissue transcriptional signature for HPV+ tumours highlighted upregulation of genes reflective of replication fork stalling as well as the cytosine deaminase *APOBEC3B* as a feature of these tumours.

Leveraging on prior work on APOBEC as a mutator in human cancers, I tested statistical measures for measuring APOBEC-mediated mutagenesis in tumour exomes, and showed that these patterns contributed to a large fraction of mutations in HPV+ tumours relative to HPV- tumours while not being part of a general mutator phenotype in viral cancers, both when assessed exome-wide and in mutations likely to be drivers of malignancy.

The serendipitous discovery of *PIK3CA* as a gene recurrently mutated by APOBEC, albeit in a hotspot specific manner lead to the identification of likely activity of APOBEC through substrate nucleotide sequences as the mutagenic process responsible for the origins of these mutations.

This relationship between a preference for helical hotspot mutations in *PIK3CA* and levels of APOBEC-activity is preserved across cancer types, and given what appears to be similar selective impact conferred by the two different *PIK3CA* hotspot mutations, establishes mutational processes as one of the key determinants of what mutations may arise in a tumour type, which will be followed by natural selection acting on those variants. To summarize, the transcriptional signature mediates the involvement of HPV in the genomic evolution of these tumours in previously undiscovered ways.

A distinct, yet complex, set of epigenetic changes defines HPV-driven tumourigenesis.

The fragmented nature of understanding regarding HPV-driven tumourigenesis that was true of transcriptome studies was also true of the methylome. Using the largest assembled set of methylomes from HPV+ tumours and corresponding normal and HPV- controls to date, I showed the existence of conserved methylation profiles and identified that only a subset of these displayed association with gene expression in canonically defined ways.

On a global scale, HPV+ tumours display a signature of genome-wide hypermethylation relative to HPV- tumours/normal tissues, which has been subject to speculation following prior studies. Some prominent HPV-induced transcriptional changes, such as the expression of the meiotic cohesin component *SYCP2* appear to be mediated by hypomethylation events at a CpG Island DMR that is methylated in HPV- tissue and normal cells.

While the contributions of methylation to shaping transcription in these tumours are fragmented and methylation by itself does not dysregulate entire pathways, some of the genes silenced by methylation, such as *PITX2*, have been shown to have potent anti-HPV functions, conferring a putative driver status for this epigenetic event.

In addition to traditionally well-known mechanisms of action for DNA methylation, my analyses uncovered a role for enhancer-element methylation events as transcriptional regulators in these tumours. While the 450k array is limited in terms of coverage of CpG sites at putative enhancers, decreases in costs of large-scale bisulfite sequencing and emerging array-based platforms that also include FANTOM5 documented enhancers (Moran, Arribas et al. 2016) should facilitate investigations into what long-range regulatory changes influence the transcriptional landscape and the evolution of HPV+ tumours.

DNA demethylating agents have recently gained attention for clinical use at low doses, with novel mechanisms currently being reported as being responsible for durable responses (Roulois, Loo Yau et al. 2015). Interfering with DNA methylation in HPV-driven tumours, particularly in context of the global hypermethylation that defines these tumours, should exert pleiotropic effects on tumour biology. Possible bases for the activity of DNA methylation inhibitors in this disease setting include interfering with E6/E7 function and stimulating anti-tumour immune responses.

Epigenomic and Transcriptional Analyses of Cell-of-origin signatures has implications for putative pathways to malignancies at different anatomical sites.

The discovery of variations in *KRT7* methylation and expression amongst different subgroups of HPV-driven tumours raised intriguing questions about the origin and evolution of HPV-driven cancers. The identification of two distinct clusters of expression of a gene set attributed to squamocolumnar junction cells is consistent with multiple models by which these tumours originate; these differences could either result from different HPV types infecting and transforming different cell types, with transcriptional heterogeneity a consequence of transcriptional variation by cell type, or from HPV and the subsequent genomic milieu in transformed cells interacting to perturb the expression of these genes to different extents in different HPV-driven cancers.

Unifying molecular profiles may improve HPV-status detection.

In Chapter 5, the datasets used for methylation analysis used a wide range of methods to call HPV-types. Traditional DNA-based methods may have issues with false-positives picked up from passenger infections or inactivated viral genome remnants, and indirect markers such as p16 staining are known to be poor indicators of HPV-status outside the Oropharynx insofar HNSC is concerned (Lassen, Primdahl et al. 2014). Approaches to call HPV-status based on RNA-seq to date are often dataset specific (Ojesina, Lichtenstein et al. 2014, 2015) (TCGA 2015) in terms of using bimodal distributions to call samples positive and there is no gold-standard for estimating if a tumour is HPV+ based on RNA-seq.

Another possibility is that there may be bystander HPV-infections or issues with contamination that compromise detection using HPV-transcription. I propose that the pan-tissue HPV-expression and methylation signatures established in this thesis may be a starting point for robust biomarkers to identify tumours that are truly driven by HPV.

Molecular heterogeneity links taxonomic variation in HPV to clinical behaviour in Cervical Cancers.

In Head and Neck Cancers, the vast majority of HPV+ tumours are caused by HPV16, with minor contributions from other types and almost never HPV types such as HPV45 and HPV18 that are vastly more common in CESC. This meant that HPV16 was likely to introduce a type specific bias when carrying out the initial analyses that defined pan-tissue signatures for HPV-driven transformation. Given prior evidence hinting at molecular and clinical differences by HPV-type, and what appeared to be different tissue-tropism amongst HPV types, I was led to explore the relationship between taxonomic groups of HPV and molecular and clinical differences.

Clinical data analysis suggested that the most pronounced clinical differences emerged when comparing HPV16-driven tumours with HPV45-driven tumours. Computational analyses of expression and methylation data from these led to the identification of large-scale differences in gene-expression and expression-associated DNA methylation patterns. Pathway analysis of the expression signature suggested these tumours displayed a profile for aggressive behaviour, even when tumours were pathologically regarded as Stage I or Stage II.

Using Random Forests for joint-clustering of data types, I showed these patterns clustered tumours into HPV45-like and HPV16-like subgroups with markedly different outcomes and defined a classifier for application in patient stratification. These aggressiveness-associated clusters also show strong overlaps with cell-types defined by cell-of-origin signature gene expression.

The analyses in this chapter suggest that prior analyses that have treated HPV driven cancers as a monolith have not recognised the existence of these functionally important differences. This also has bearings on interpretation of gene-expression changes that are found in pan-tissue signatures as “HPV-specific” phenomena instead of “HPV-subtype specific” phenomena. For instance, HPV45+ tumours display high levels of CCND1 expression relative to HPV16+ tumours and counter the notion that CCND1 activation mediates cell death if *CDKN2A* expression is silenced across all HPV-driven tumours instead of just HPV16-driven tumours (McLaughlin-Drubin, Park et al. 2013).

It is particularly striking that clustering using the HPV45 signature has prognostic value across the entirety of the cohort, independent of HPV type, and has prognostic value in cervical cancers caused by HPV16 and HPV18, making it possible to apply these patterns for patient stratification across all cervical cancers when samples are initially biopsied. The fact that these Aggressiveness Classes cross lines of cellular origin and histology identifies fundamental truths about the behaviour of these tumours. There is an opportunity for future work to build upon these findings to further understand the determinants of these transcriptional and epigenetic profiles, be they viral or cellular, and to test this classification scheme for patient management.

Analysis of Immune Microenvironment across HPV+ cancers identifies two broad immune profiles.

Finally, I developed regression based methods for estimating the immune microenvironment of HPV+ tumours, and found that HPV45-like tumours were marked by depleted levels of CD8+ TILs and high levels of infiltrating Natural Killer and Granulocytes, consistent with prior evidence for patterns of immune-cells and prognosis, which is also consistent with high levels of expression of *CD276*, a checkpoint-regulator, in these tumours, suggesting a potential target for immunotherapy (Pardoll 2012). Joint clustering based on immune effectors and infiltration estimates identified an immunoreactive and an immunodepleted group within HPV+ tumours, with the former defined by overexpression of multiple mediators of immune cell signalling and function. The presence of high levels of immune infiltrates and checkpoint activity within these tumours indicates potential for the use of immunotherapies in their treatment. A considerable fraction of HPV45-like tumours also display an immunoenriched subtype characterised by high levels of CD8+ lymphocytes and Tregs, identifying a subgroup that may benefit from immune checkpoint modulation.

The *MethylCIBERSORT* approach*,* developed for deep deconvolution of immune cell subsets using methylation data, will facilitate more detailed future investigations into the interplay between molecular alterations and immune cell distributions across cancer types. In the future, higher resolution DNA methylation assays should permit greater discrimination of a larger number of cell types, and refine the profiles used to estimate cellular composition by using immune cells obtained through the direct dissociation of tumour-associated immune cells and other infiltrating cells.

To summarise, these findings implicate marked differences in the means of immunoevasion amongst HPV+ tumours and lay the foundations for further work into manipulating these pathways for therapeutic purposes.

Preservation of molecular profiles has implications for therapy.

Results from Chapter 5 demonstrate that epigenetic profiles and the expression profiles of DNA-methylation associated genes are preserved across the tumour cohort, including genes that are part of the metasignature and those that are putative HPV-associated drivers. This implies that HPV plays a continual role in maintaining the viability of these tumours, which has been confirmed by others in cell-line studies involving the knockdown of E6/E7 (Rampias, Sasaki et al. 2009, Chang, Kuo et al. 2010).

This dependence on HPV allows for the development of customised T-cell therapies to target E6/E7 antigen-displaying cells for immune destruction, and small-scale trials have already reported some successes using these approaches (Stevanovic, Draper et al. 2015) . The presence of HPV, and continued expression thereof, in every cell (NB – clonal neoantigens have been proposed to be associated with more potent anti-tumour immune activity (McGranahan, Furness et al. 2016)), in combination with general immunomodulatory strategies to convert the immune microenvironment of HPV+ tumours into one that is amenable to tumour rejection is a feasible pathway towards cures. One outstanding question that has not been addressed in the thesis, however, is whether metastatic tumours may display vastly different molecular profiles having undergone, in the course of tumour evolution, alterations that subvert the signatures seen in primary tumours.

Final Synthesis

Finally, I wish to integrate the aforementioned findings into a unified model that links origin, molecular profiles, and outcomes in HPV+ tumours, and raises questions for further research. Fundamentally, I propose that different high-risk HPV-types show differential affinity for transforming two different cell types, one involving homologous populations to the squamocolumnar junction cells of the cervix, and one a different population for which HPV16 shows greater affinity than other cell types.

This is consistent with the varying tropism for different HPV types to different regions. HPV18, for instance, is restricted to the anus and the uterine cervix, both of which show junction-cell populations, whereas in the penis and head and neck regions, HPV18 and HPV45 are seldom seen.

Despite differences in the cell-of-origin for these tumours, the activity of HPV-oncoproteins can induce a common set of transcriptional, epigenetic and proteomic changes. These can have consequences including genomic instability and replication fork stalling which potentiate the off-target activity of APOBEC3. APOBEC3-mediated mutagenesis then becomes the primary mechanism through which genomic disruption occurs on the pathway to full-blown malignancy.

Unified DNA methylation patterns may be associated with candidate driver epigenetic changes that indicate epigenetic therapy may exert pleiotropic effects that ultimately subvert HPV-oncoprotein function and HPV-driven cancer cells to collapse.

Despite these common changes, either through the activity of HPV-oncoproteins themselves, or the underlying transcriptional and epigenomic architectures of the cell-of-origin, additional heterogeneity is induced. Some of these involve marked changes in pathways associated with metastatic behaviour and a pro-inflammatory microenvironment, that translates to abysmal clinical outlook within a subset of tumours. Finally, the immune microenvironment appears to be a key selective pressure shaping the evolution of these tumours, and two distinct patterns of immune infiltration and activity, with associated molecular correlates, are evident.

In toto, the work in this thesis serves to help identify if a tumour is HPV-driven using transcription/DNA-methylation markers, links patterns that unify HPV-driven cancers across tissues to mutagenesis and transformation, offering potential scope for prevention, and then identifies tumours likely to have diverse cellular origins and clinical behaviour. It finally identifies markers of immune activity that may be employed to judge who will benefit from immunotherapies, especially if they display molecular features of the aggressive subtype, in the process confirming each of the three core hypotheses laid out at the outset.

References

Agnihotri, S., A. Wolf, D. M. Munoz, C. J. Smith, A. Gajadhar, A. Restrepo, I. D. Clarke, G. N. Fuller, S. Kesari, P. B. Dirks, C. J. McGlade, W. L. Stanford, K. Aldape, P. S. Mischel, C. Hawkins and A. Guha (2011). "A GATA4-regulated tumor suppressor network represses formation of malignant human astrocytomas." J Exp Med **208**(4): 689-702.

Agrawal, N., M. J. Frederick, C. R. Pickering, C. Bettegowda, K. Chang, R. J. Li, C. Fakhry, T. X. Xie, J. Zhang, J. Wang, N. Zhang, A. K. El-Naggar, S. A. Jasser, J. N. Weinstein, L. Trevino, J. A. Drummond, D. M. Muzny, Y. Wu, L. D. Wood, R. H. Hruban, W. H. Westra, W. M. Koch, J. A. Califano, R. A. Gibbs, D. Sidransky, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, D. A. Wheeler, K. W. Kinzler and J. N. Myers (2011). "Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1." Science **333**(6046): 1154-1157.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell and M. R. Stratton (2013). "Signatures of mutational processes in human cancer." Nature **500**(7463): 415-421.

Anderson, A. C. (2014). "Tim-3: an emerging target in the cancer immunotherapy landscape." Cancer Immunol Res **2**(5): 393-398.

Ang, K. K., J. Harris, R. Wheeler, R. Weber, D. I. Rosenthal, P. F. Nguyen-Tan, W. H. Westra, C. H. Chung, R. C. Jordan, C. Lu, H. Kim, R. Axelrod, C. C. Silverman, K. P. Redmond and M. L. Gillison (2010). "Human papillomavirus and survival of patients with oropharyngeal cancer." N Engl J Med **363**(1): 24-35.

Aran, D., M. Sirota and A. J. Butte (2015). "Systematic pan-cancer analysis of tumour purity." Nat Commun **6**: 8971.

Aran, D., M. Sirota and A. J. Butte (2016). "Corrigendum: Systematic pan-cancer analysis of tumour purity." Nat Commun **7**: 10707.

Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry (2014). "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays." Bioinformatics **30**(10): 1363-1369.

Bao, X., J. Tang, V. Lopez-Pajares, S. Tao, K. Qu, G. R. Crabtree and P. A. Khavari (2013). "ACTL6a enforces the epidermal progenitor state by suppressing SWI/SNF-dependent induction of KLF4." Cell Stem Cell **12**(2): 193-203.

Baussano, I., F. Lazzarato, G. Ronco, J. Dillner and S. Franceschi (2013). "Benefits of catch-up in vaccination against human papillomavirus in medium- and low-income countries." Int J Cancer **133**(8): 1876-1881.

Benjamini, Y. (2010). "Discovering the false discovery rate." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(4): 405-416.

Benjamini, Y., A. M. Krieger and D. Yekutieli (2006). "Adaptive linear step-up procedures that control the false discovery rate." Biometrika **93**(3): 491-507.

Bernard, H. U., R. D. Burk, Z. Chen, K. van Doorslaer, H. zur Hausen and E. M. de Villiers (2010). "Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments." Virology **401**(1): 70-79.

Blattler, A., L. Yao, H. Witt, Y. Guo, C. M. Nicolet, B. P. Berman and P. J. Farnham (2014). "Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes." Genome Biol **15**(9): 469.

Bold, T. D. and J. D. Ernst (2012). "CD4+ T cell-dependent IFN-gamma production by CD8+ effector T cells in Mycobacterium tuberculosis infection." J Immunol **189**(5): 2530-2536.

Bouvard, V., R. Baan, K. Straif, Y. Grosse, B. Secretan, F. El Ghissassi, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet and V. Cogliano (2009). "A review of human carcinogens--Part B: biological agents." Lancet Oncol **10**(4): 321-322.

Bouvard, V., R. Baan, K. Straif, Y. Grosse, B. Secretan, F. E. Ghissassi, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet and V. Cogliano (2014). "A review of human carcinogens 2014;Part B: biological agents." The Lancet Oncology **10**(4): 321-322.

Brasa, S., A. Mueller, S. Jacquemont, F. Hahne, I. Rozenberg, T. Peters, Y. He, C. McCormack, F. Gasparini, S. D. Chibout, O. Grenet, J. Moggs, B. Gomez-Mancilla and R. Terranova (2016). "Reciprocal changes in DNA methylation and hydroxymethylation and a broad repressive epigenetic switch characterize FMR1 transcriptional silencing in fragile X syndrome." Clin Epigenetics **8**: 15.

Brocks, D., Y. Assenov, S. Minner, O. Bogatyrova, R. Simon, C. Koop, C. Oakes, M. Zucknick, D. B. Lipka, J. Weischenfeldt, L. Feuerbach, R. Cowper-Sal Lari, M. Lupien, B. Brors, J. Korbel, T. Schlomm, A. Tanay, G. Sauter, C. Gerhauser and C. Plass (2014). "Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer." Cell Rep **8**(3): 798-806.

Brosh, R. and V. Rotter (2009). "When mutants gain new powers: news from the mutant p53 field." Nat Rev Cancer **9**(10): 701-713.

Buitrago-Perez, A., G. Garaulet, A. Vazquez-Carballo, J. M. Paramio and R. Garcia-Escudero (2009). "Molecular Signature of HPV-Induced Carcinogenesis: pRb, p53 and Gene Expression Profiling." Curr Genomics **10**(1): 26-34.

Burger, R. A., B. J. Monk, T. Kurosaki, H. Anton-Culver, S. A. Vasilev, M. L. Berman and S. P. Wilczynski (1996). "Human papillomavirus type 18: association with poor prognosis in early stage cervical cancer." J Natl Cancer Inst **88**(19): 1361-1368.

Burns, M. B., L. Lackey, M. A. Carpenter, A. Rathore, A. M. Land, B. Leonard, E. W. Refsland, D. Kotandeniya, N. Tretyakova, J. B. Nikas, D. Yee, N. A. Temiz, D. E. Donohue, R. M. McDougle, W. L. Brown, E. K. Law and R. S. Harris (2013). "APOBEC3B is an enzymatic source of mutation in breast cancer." Nature **494**(7437): 366-370.

Burns, M. B., N. A. Temiz and R. S. Harris (2013). "Evidence for APOBEC3B mutagenesis in multiple human cancers." Nat Genet **45**(9): 977-983.

Butcher, L. M. and S. Beck (2015). "Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data." Methods **72**: 21-28.

Carter, S. L., K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson and G. Getz (2012). "Absolute quantification of somatic DNA alterations in human cancer." Nat Biotechnol **30**(5): 413-421.

Ceccarelli, M., F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, S. Anjum, J. Wang, G. Manyam, P. Zoppoli, S. Ling, A.

A. Rao, M. Grifford, A. D. Cherniack, H. Zhang, L. Poisson, C. G. Carlotti, Jr., D. P. Tirapelli, A. Rao, T. Mikkelsen, C. C. Lau, W. K. Yung, R. Rabadan, J. Huse, D. J. Brat, N. L. Lehman, J. S. Barnholtz-Sloan, S. Zheng, K. Hess, G. Rao, M. Meyerson, R. Beroukhim, L. Cooper, R. Akbani, M. Wrensch, D. Haussler, K. D. Aldape, P. W. Laird, D. H. Gutmann, H. Noushmehr, A. Iavarone and R. G. Verhaak (2016). "Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma." Cell **164**(3): 550-563.

Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander and N. Schultz (2012). "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data." Cancer Discov **2**(5): 401-404.

Chakravarthy, A., S. Henderson, S. M. Thirdborough, C. H. Ottensmeier, X. Su, M. Lechner, A. Feber, G. J. Thomas and T. R. Fenton (2016). "Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx." Journal of Clinical Oncology: JCO682955.

Chakravarthy, A., S. Henderson, S. M. Thirdborough, C. H. Ottensmeier, X. Su, M. Lechner, A. Feber, G. J. Thomas and T. R. Fenton (2016). "Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx." J Clin Oncol **34**(34): 4132-4141.

Chang, J. T., T. F. Kuo, Y. J. Chen, C. C. Chiu, Y. C. Lu, H. F. Li, C. R. Shen and A. J. Cheng (2010). "Highly potent and specific siRNAs against E6 or E7 genes of HPV16- or HPV18-infected cervical cancers." Cancer Gene Ther **17**(12): 827-836.

Chen, D., T. Cui, W. E. Ek, H. Liu, H. Wang and U. Gyllensten (2015). "Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability." Carcinogenesis **36**(9): 992-998.

Chen, J., B. F. Miller and A. V. Furano (2014). "Repair of naturally occurring mismatches can induce mutations in flanking DNA." Elife **3**: e02001.

Chen, X., J. Miao, H. Wang, F. Zhao, J. Hu, P. Gao, Y. Wang, L. Zhang and M. Yan (2015). "The anti-inflammatory activities of Ainsliaea fragrans Champ. extract and its components in lipopolysaccharide-stimulated RAW264.7 macrophages through inhibition of NF-kappaB pathway." J Ethnopharmacol **170**: 72-80.

Chowdhury, U. R., R. S. Samant, O. Fodstad and L. A. Shevde (2009). "Emerging role of nuclear protein 1 (NUPR1) in cancer biology." Cancer Metastasis Rev **28**(1-2): 225-232.

Chung, C. H., Q. Zhang, C. S. Kong, J. Harris, E. J. Fertig, P. M. Harari, D. Wang, K. P. Redmond, G. Shenouda, A. Trotti, D. Raben, M. L. Gillison, R. C. Jordan and Q. T. Le (2014). "p16 protein expression and human papillomavirus status as prognostic biomarkers of nonoropharyngeal head and neck squamous cell carcinoma." J Clin Oncol **32**(35): 3930-3938.

Clark, D. W., A. Mitra, R. A. Fillmore, W. G. Jiang, R. S. Samant, O. Fodstad and L. A. Shevde (2008). "NUPR1 interacts with p53, transcriptionally regulates p21 and rescues breast epithelial cells from doxorubicin-induced genotoxic stress." Curr Cancer Drug Targets **8**(5): 421-430.

Collins, S. I., C. Constandinou-Williams, K. Wen, L. S. Young, S. Roberts, P. G. Murray and C. B. Woodman (2009). "Disruption of the E2 gene is a common and early event in the natural history of cervical human papillomavirus infection: a longitudinal cohort study." Cancer Res **69**(9): 3828-3832.

Davies, R. G., K. M. Wagstaff, E. A. McLaughlin, K. L. Loveland and D. A. Jans (2013). "The BRCA1-binding protein BRAP2 can act as a cytoplasmic retention factor for nuclear and nuclear envelope-localizing testicular proteins." Biochim Biophys Acta **1833**(12): 3436-3444.

Day, T. and C. Vaziri (2009). "HPV E6 oncoprotein prevents recovery of stalled replication forks independently of p53 degradation." Cell Cycle **8**(14): 2138.

De Carvalho, D. D., S. Sharma, J. S. You, S. F. Su, P. C. Taberlay, T. K. Kelly, X. Yang, G. Liang and P. A. Jones (2012). "DNA methylation screening identifies driver epigenetic events of cancer cell survival." Cancer Cell **21**(5): 655-667.

DeCaprio, J. A. (2014). "Human papillomavirus type 16 E7 perturbs DREAM to promote cellular proliferation and mitotic gene expression." Oncogene **33**(31): 4036-4038.

Dowhanick, J. J., A. A. McBride and P. M. Howley (1995). "Suppression of cellular proliferation by the papillomavirus E2 protein." J Virol **69**(12): 7791-7799.

Ekholm-Reed, S., J. Mendez, D. Tedesco, A. Zetterberg, B. Stillman and S. I. Reed (2004). "Deregulation of cyclin E in human cells interferes with prereplication complex assembly." J Cell Biol **165**(6): 789-800.

ENCODE (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Fahey, L. M., A. B. Raff, D. M. Da Silva and W. M. Kast (2009). "Reversal of human papillomavirus-specific T cell immune suppression through TLR agonist treatment of Langerhans cells exposed to human papillomavirus type 16." J Immunol **182**(5): 2919-2928.

Fakhry, C., W. H. Westra, S. Li, A. Cmelak, J. A. Ridge, H. Pinto, A. Forastiere and M. L. Gillison (2008). "Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial." J Natl Cancer Inst **100**(4): 261-269.

Farkas, S. A., N. Milutin-Gasperov, M. Grce and T. K. Nilsson (2013). "Genome-wide DNA methylation assay reveals novel candidate biomarker genes in cervical cancer." Epigenetics **8**(11): 1213-1225.

Feber, A., M. Arya, P. de Winter, M. Saqib, R. Nigam, P. R. Malone, W. S. Tan, S. Rodney, M. Lechner, A. Freeman, C. Jameson, A. Muneer, S. Beck and J. D. Kelly (2015). "Epigenetics markers of metastasis and HPV-induced tumorigenesis in penile cancer." Clin Cancer Res **21**(5): 1196-1206.

Ferlay, J., I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman and F. Bray (2014). "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012." Int J Cancer.

Flavahan, W. A., Y. Drier, B. B. Liau, S. M. Gillespie, A. S. Venteicher, A. O. Stemmer-Rachamimov, M. L. Suva and B. E. Bernstein (2016). "Insulator dysfunction and oncogene activation in IDH mutant gliomas." Nature **529**(7584): 110-114.

Funes, J. M., M. Quintero, S. Henderson, D. Martinez, U. Qureshi, C. Westwood, M. O. Clements, D. Bourboulia, R. B. Pedley, S. Moncada and C. Boshoff (2007). "Transformation of human mesenchymal stem cells increases their dependency on oxidative phosphorylation for energy production." Proc Natl Acad Sci U S A **104**(15): 6223-6228.

Gautier, L., L. Cope, B. M. Bolstad and R. A. Irizarry (2004). "affy--analysis of Affymetrix GeneChip data at the probe level." Bioinformatics **20**(3): 307-315.

Gilbert, D. C., E. Serup-Hansen, D. Linnemann, E. Hogdall, C. Bailey, J. Summers, H. Havsteen and G. J. Thomas (2016). "Tumour-infiltrating lymphocyte scores effectively stratify outcomes over and above p16 post chemo-radiotherapy in anal cancer." Br J Cancer.

Gillison, M. L., A. K. Chaturvedi, W. F. Anderson and C. Fakhry (2015). "Epidemiology of Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma." J Clin Oncol **33**(29): 3235-3242.

Gonzalez-Perez, A., A. Jene-Sanz and N. Lopez-Bigas (2013). "The mutational landscape of chromatin regulatory factors across 4,623 tumor samples." Genome Biol **14**(9): r106.

Gray, E., M. R. Pett, D. Ward, D. M. Winder, M. A. Stanley, I. Roberts, C. G. Scarpini and N. Coleman (2010). "In vitro progression of human papillomavirus 16 episome-associated cervical neoplasia displays fundamental similarities to integrant-associated carcinogenesis." Cancer Res **70**(10): 4081-4091.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Handler, N. S., M. Z. Handler, S. Majewski and R. A. Schwartz (2015). "Human papillomavirus vaccine trials and tribulations: Vaccine efficacy." J Am Acad Dermatol **73**(5): 759-767; quiz 767-758.

Harbour, J. W. and D. C. Dean (2000). "The Rb/E2F pathway: expanding roles and emerging paradigms." Genes Dev **14**(19): 2393-2409.

He, H. and Y. Luo (2012). "Brg1 regulates the transcription of human papillomavirus type 18 E6 and E7 genes." Cell Cycle **11**(3): 617-627.

Henderson, S., A. Chakravarthy, X. Su, C. Boshoff and T. R. Fenton (2014). "APOBEC-Mediated Cytosine Deamination Links PIK3CA Helical Domain Mutations to Human Papillomavirus-Driven Tumor Development." Cell Reports **7**(6): 1833-1841.

Henderson, S., A. Chakravarthy, X. Su, C. Boshoff and T. R. Fenton (2014). "APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development." Cell Rep **7**(6): 1833-1841.

Herfs, M., Y. Yamamoto, A. Laury, X. Wang, M. R. Nucci, M. E. McLaughlin-Drubin, K. Munger, S. Feldman, F. D. McKeon, W. Xian and C. P. Crum (2012). "A discrete population of squamocolumnar junction cells implicated in the pathogenesis of cervical cancer." Proc Natl Acad Sci U S A **109**(26): 10516-10521.

Hong, S., K. P. Mehta and L. A. Laimins (2011). "Suppression of STAT-1 expression by human papillomaviruses is necessary for differentiation-dependent genome amplification and plasmid maintenance." J Virol **85**(18): 9486-9494.

Houseman, E. A., J. Molitor and C. J. Marsit (2014). "Reference-free cell mixture adjustments in analysis of DNA methylation data." Bioinformatics **30**(10): 1431-1439.

Hovestadt, V., D. T. Jones, S. Picelli, W. Wang, M. Kool, P. A. Northcott, M. Sultan, K. Stachurski, M. Ryzhova, H. J. Warnatz, M. Ralser, S. Brun, J. Bunt, N. Jager, K. Kleinheinz, S. Erkek, U. D. Weber, C. C. Bartholomae, C. von Kalle, C. Lawerenz, J. Eils, J. Koster, R. Versteeg, T. Milde, O. Witt, S. Schmidt, S. Wolf, T. Pietsch, S. Rutkowski, W. Scheurlen, M. D. Taylor, B. Brors, J. Felsberg, G. Reifenberger, A. Borkhardt, H. Lehrach, R. J. Wechsler-Reya, R. Eils, M. L. Yaspo, P. Landgraf, A. Korshunov, M. Zapatka, B. Radlwimmer, S. M. Pfister and P. Lichter (2014). "Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing." Nature **510**(7506): 537-541.

Huang, C. H., D. Mandelker, O. Schmidt-Kittler, Y. Samuels, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, S. B. Gabelli and L. M. Amzel (2007). "The structure of a human p110alpha/p85alpha complex elucidates the effects of oncogenic PI3Kalpha mutations." Science **318**(5857): 1744-1748.

Hugo, W., J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, E. Seja, S. Lomeli, X. Kong, M. C. Kelley, J. A. Sosman, D. B. Johnson, A.

Ribas and R. S. Lo (2016). "Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma." Cell **165**(1): 35-44.

Im, S. S., S. P. Wilczynski, R. A. Burger and B. J. Monk (2003). "Early stage cervical cancers containing human papillomavirus type 18 DNA have more nodal metastasis and deeper stromal invasion." Clin Cancer Res **9**(11): 4145-4150.

Jaffe, A. E. and R. A. Irizarry (2014). "Accounting for cellular heterogeneity is critical in epigenome-wide association studies." Genome Biol **15**(2): R31.

Janas, M. L., P. Groves, N. Kienzle and A. Kelso (2005). "IL-2 regulates perforin and granzyme gene expression in CD8+ T cells independently of its effects on survival and proliferation." J Immunol **175**(12): 8003-8010.

Kaczkowski, B., M. Morevati, M. Rossing, F. Cilius and B. Norrild (2012). "A Decade of Global mRNA and miRNA Profiling of HPV-Positive Cell Lines and Clinical Specimens." Open Virol J **6**: 216-231.

Kadoch, C., D. C. Hargreaves, C. Hodges, L. Elias, L. Ho, J. Ranish and G. R. Crabtree (2013). "Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy." Nat Genet **45**(6): 592-601.

Kang, S., A. G. Bader and P. K. Vogt (2005). "Phosphatidylinositol 3-kinase mutations identified in human cancer are oncogenic." Proc Natl Acad Sci U S A **102**(3): 802-807.

Kaufhold, S. and B. Bonavida (2014). "Central role of Snail1 in the regulation of EMT and resistance in cancer: a target for therapeutic intervention." J Exp Clin Cancer Res **33**: 62.

Kimple, R. J., M. A. Smith, G. C. Blitzer, A. D. Torres, J. A. Martin, R. Z. Yang, C. R. Peet, L. D. Lorenz, K. P. Nickel, A. J. Klingelhutz, P. F. Lambert and P. M. Harari (2013). "Enhanced radiation sensitivity in HPV-positive head and neck cancer." Cancer Res **73**(15): 4791-4800.

Klaes, R., T. Friedrich, D. Spitkovsky, R. Ridder, W. Rudy, U. Petry, G. Dallenbach-Hellweg, D. Schmidt and M. von Knebel Doeberitz (2001). "Overexpression of p16(INK4A) as a specific marker for dysplastic and neoplastic epithelial cells of the cervix uteri." Int J Cancer **92**(2): 276-284.

Koestler, D. C., B. Christensen, M. R. Karagas, C. J. Marsit, S. M. Langevin, K. T. Kelsey, J. K. Wiencke and E. A. Houseman (2013). "Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis." Epigenetics **8**(8): 816-826.

Kravchenko-Balasha, N., S. Mizrachy-Schwartz, S. Klein and A. Levitzki (2009). "Shift from apoptotic to necrotic cell death during human papillomavirus-induced transformation of keratinocytes." J Biol Chem **284**(17): 11717-11727.

Kreimer, A. R., P. Brennan, K. A. Lang Kuhs, T. Waterboer, G. Clifford, S. Franceschi, A. Michel, M. Willhauck-Fleckenstein, E. Riboli, X. Castellsague, A. Hildesheim, R. T. Fortner, R. Kaaks, D. Palli, I. Ljuslinder, S. Panico, F. Clavel-Chapelon, M. C. Boutron-Ruault, S. Mesrine, A. Trichopoulou, P. Lagiou, D. Trichopoulos, P. H. Peeters, A. J. Cross, H. B. Bueno-de-Mesquita, P. Vineis, N. Larranaga, V. Pala, M. J. Sanchez, C. Navarro, A. Barricarte, R. Tumino, K. T. Khaw, N. Wareham, H. Boeing, A. Steffen, R. C. Travis, J. R. Quiros, E. Weiderpass, M. Pawlita and M. Johansson (2015). "Human papillomavirus antibodies and future risk of anogenital cancer: a nested case-control study in the European prospective investigation into cancer and nutrition study." J Clin Oncol **33**(8): 877-884.

Landau, D. A., K. Clement, M. J. Ziller, P. Boyle, J. Fan, H. Gu, K. Stevenson, C. Sougnez, L. Wang, S. Li, D. Kotliar, W. Zhang, M. Ghandi, L. Garraway, S. M. Fernandes, K. J. Livak, S. Gabriel, A. Gnirke, E. S. Lander, J. R. Brown, D. Neuberg, P. V. Kharchenko, N. Hacohen, G. Getz, A. Meissner

and C. J. Wu (2014). "Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia." Cancer Cell **26**(6): 813-825.

Lassen, P., H. Primdahl, J. Johansen, C. A. Kristensen, E. Andersen, L. J. Andersen, J. F. Evensen, J. G. Eriksen and J. Overgaard (2014). "Impact of HPV-associated p16-expression on radiotherapy outcome in advanced oropharynx and non-oropharynx cancer." Radiother Oncol **113**(3): 310-316.

Law, C. W., Y. Chen, W. Shi and G. K. Smyth (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." Genome Biol **15**(2): R29.

Lechner, M., T. Fenton, J. West, G. Wilson, A. Feber, S. Henderson, C. Thirlwell, H. K. Dibra, A. Jay, L. Butcher, A. R. Chakravarthy, F. Gratrix, N. Patel, F. Vaz, P. O'Flynn, N. Kalavrezos, A. E. Teschendorff, C. Boshoff and S. Beck (2013). "Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma." Genome Med **5**(2): 15.

Lechner, M., G. M. Frampton, T. Fenton, A. Feber, G. Palmer, A. Jay, N. Pillay, M. Forster, M. T. Cronin, D. Lipson, V. A. Miller, T. A. Brennan, S. Henderson, F. Vaz, P. O'Flynn, N. Kalavrezos, R. Yelensky, S. Beck, P. J. Stephens and C. Boshoff (2013). "Targeted next-generation sequencing of head and neck squamous cell carcinoma identifies novel genetic alterations in HPV+ and HPV- tumors." Genome Med **5**(5): 49.

Lee, K., A. Y. Lee, Y. K. Kwon and H. Kwon (2011). "Suppression of HPV E6 and E7 expression by BAF53 depletion in cervical cancer cells." Biochem Biophys Res Commun **412**(2): 328-333.

Lee, S. T., Z. Li, Z. Wu, M. Aau, P. Guan, R. K. Karuturi, Y. C. Liou and Q. Yu (2011). "Context-specific regulation of NF-kappaB target gene expression by EZH2 in breast cancers." Mol Cell **43**(5): 798-810.

Lee, Y. Y., C. H. Choi, H. J. Kim, T. J. Kim, J. W. Lee, J. H. Lee, D. S. Bae and B. G. Kim (2012). "Pretreatment neutrophil:lymphocyte ratio as a prognostic factor in cervical carcinoma." Anticancer Res **32**(4): 1555-1561.

Lewis, P. W., M. M. Muller, M. S. Koletsky, F. Cordero, S. Lin, L. A. Banaszynski, B. A. Garcia, T. W. Muir, O. J. Becher and C. D. Allis (2013). "Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma." Science **340**(6134): 857-861.

Liu, Z. and T. M. Roberts (2006). "Human tumor mutants in the p110alpha subunit of PI3K." Cell Cycle **5**(7): 675-677.

Lizio, M., J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C. J. Mungall, E. Arner, J. K. Baillie, N. Bertin, H. Bono, M. de Hoon, A. D. Diehl, E. Dimont, T. C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T. F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E. A. Schultes, P. A. t Hoen, Z. Tatum, M. Thompson, T. Toyoda, D. W. Wright, C. O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A. R. Forrest and H. Kawaji (2015). "Gateways to the FANTOM5 promoter level mammalian expression atlas." Genome Biol **16**: 22.

Llorens, C., G. P. Bernet, S. Ramasamy, C. Feschotte and A. Moya (2012). "On the transposon origins of mammalian SCAND3 and KRBA2, two zinc-finger genes carrying an integrase/transposase domain." Mob Genet Elements **2**(5): 205-210.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Lyford-Pike, S., S. Peng, G. D. Young, J. M. Taube, W. H. Westra, B. Akpeng, T. C. Bruno, J. D. Richmon, H. Wang, J. A. Bishop, L. Chen, C. G. Drake, S. L. Topalian, D. M. Pardoll and S. I. Pai (2013). "Evidence for a role of the PD-1:PD-L1 pathway in immune resistance of HPV-associated head and neck squamous cell carcinoma." Cancer Res **73**(6): 1733-1741.

Mack, S. C., H. Witt, R. M. Piro, L. Gu, S. Zuyderduyn, A. M. Stutz, X. Wang, M. Gallo, L. Garzia, K. Zayne, X. Zhang, V. Ramaswamy, N. Jager, D. T. Jones, M. Sill, T. J. Pugh, M. Ryzhova, K. M. Wani, D. J. Shih, R. Head, M. Remke, S. D. Bailey, T. Zichner, C. C. Faria, M. Barszczyk, S. Stark, H. Seker-Cin, S. Hutter, P. Johann, S. Bender, V. Hovestadt, T. Tzaridis, A. M. Dubuc, P. A. Northcott, J. Peacock, K. C. Bertrand, S. Agnihotri, F. M. Cavalli, I. Clarke, K. Nethery-Brokx, C. L. Creasy, S. K. Verma, J. Koster, X. Wu, Y. Yao, T. Milde, P. Sin-Chan, J. Zuccaro, L. Lau, S. Pereira, P. Castelo-Branco, M. Hirst, M. A. Marra, S. S. Roberts, D. Fults, L. Massimi, Y. J. Cho, T. Van Meter, W. Grajkowska, B. Lach, A. E. Kulozik, A. von Deimling, O. Witt, S. W. Scherer, X. Fan, K. M. Muraszko, M. Kool, S. L. Pomeroy, N. Gupta, J. Phillips, A. Huang, U. Tabori, C. Hawkins, D. Malkin, P. N. Kongkham, W. A. Weiss, N. Jabado, J. T. Rutka, E. Bouffet, J. O. Korbel, M. Lupien, K. D. Aldape, G. D. Bader, R. Eils, P. Lichter, P. B. Dirks, S. M. Pfister, A. Korshunov and M. D. Taylor (2014). "Epigenomic alterations define lethal CIMP-positive ependymomas of infancy." Nature **506**(7489): 445-450.

Mailand, N., I. Gibbs-Seymour and S. Bekker-Jensen (2013). "Regulation of PCNA-protein interactions for genome stability." Nat Rev Mol Cell Biol **14**(5): 269-282.

Malagon, T., M. Drolet, M. C. Boily, E. L. Franco, M. Jit, J. Brisson and M. Brisson (2012). "Cross-protective efficacy of two human papillomavirus vaccines: a systematic review and meta-analysis." Lancet Infect Dis **12**(10): 781-789.

Man, S. M., R. Karki and T. D. Kanneganti (2016). "AIM2 inflammasome in infection, cancer, and autoimmunity: Role in DNA sensing, inflammation, and innate immunity." Eur J Immunol **46**(2): 269-280.

Mansour, M. R., B. J. Abraham, L. Anders, A. Berezovskaya, A. Gutierrez, A. D. Durbin, J. Etchin, L. Lawton, S. E. Sallan, L. B. Silverman, M. L. Loh, S. P. Hunger, T. Sanda, R. A. Young and A. T. Look (2014). "Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element." Science **346**(6215): 1373-1377.

Marot, G., J.-L. Foulley, C.-D. Mayer and F. Jaffrézic (2009). "Moderated effect size and p-value combinations for microarray meta-analyses." Bioinformatics.

Marur, S., G. D'Souza, W. H. Westra and A. A. Forastiere (2010). "HPV-associated head and neck cancer: a virus-related cancer epidemic." Lancet Oncol **11**(8): 781-789.

Maunakea, A. K., I. Chepelev, K. Cui and K. Zhao (2013). "Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition." Cell Res **23**(11): 1256-1269.

McGranahan, N., A. J. Furness, R. Rosenthal, S. Ramskov, R. Lyngaa, S. K. Saini, M. Jamal-Hanjani, G. A. Wilson, N. J. Birkbak, C. T. Hiley, T. B. Watkins, S. Shafi, N. Murugaesu, R. Mitter, A. U. Akarca, J. Linares, T. Marafioti, J. Y. Henry, E. M. Van Allen, D. Miao, B. Schilling, D. Schadendorf, L. A. Garraway, V. Makarov, N. A. Rizvi, A. Snyder, M. D. Hellmann, T. Merghoub, J. D. Wolchok, S. A. Shukla, C. J. Wu, K. S. Peggs, T. A. Chan, S. R. Hadrup, S. A. Quezada and C. Swanton (2016). "Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade." Science **351**(6280): 1463-1469.

McLaughlin-Drubin, M. E., C. P. Crum and K. Munger (2011). "Human papillomavirus E7 oncoprotein induces KDM6A and KDM6B histone demethylase expression and causes epigenetic reprogramming." Proc Natl Acad Sci U S A **108**(5): 2130-2135.

McLaughlin-Drubin, M. E., K. W. Huh and K. Munger (2008). "Human papillomavirus type 16 E7 oncoprotein associates with E2F6." J Virol **82**(17): 8695-8705.

McLaughlin-Drubin, M. E., D. Park and K. Munger (2013). "Tumor suppressor p16INK4A is necessary for survival of cervical carcinoma cell lines." Proc Natl Acad Sci U S A **110**(40): 16175-16180.

Meissl, K., S. Macho-Maschler, M. Muller and B. Strobl (2015). "The good and the bad faces of STAT1 in solid tumours." Cytokine.

Mine, K. L., N. Shulzhenko, A. Yambartsev, M. Rochman, G. F. O. Sanson, M. Lando, S. Varma, J. Skinner, N. Volfovsky, T. Deng, S. M. F. Brenna, C. R. N. Carvalho, J. C. L. Ribalta, M. Bustin, P. Matzinger, I. D. C. G. Silva, H. Lyng, M. Gerbase-DeLima and A. Morgun (2013). "Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer." Nat Commun **4**: 1806.

Mittal, D., M. M. Gubin, R. D. Schreiber and M. J. Smyth (2014). "New insights into cancer immunoediting and its three component phases--elimination, equilibrium and escape." Curr Opin Immunol **27**: 16-25.

Mizunuma, M., Y. Yokoyama, M. Futagami, M. Aoki, Y. Takai and H. Mizunuma (2015). "The pretreatment neutrophil-to-lymphocyte ratio predicts therapeutic response to radiation therapy and concurrent chemoradiation therapy in uterine cervical cancer." Int J Clin Oncol **20**(5): 989-996.

Moogk, D., S. Zhong and Z. Yu (2016). "Constitutive Lck Activity Drives Sensitivity Differences between CD8+ Memory T Cell Subsets." **197**(2): 644-654.

Mora, A., G. K. Sandve, O. S. Gabrielsen and R. Eskeland (2015). "In the loop: promoter-enhancer interactions and bioinformatics." Brief Bioinform.

Moran, S., C. Arribas and M. Esteller (2016). "Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences." Epigenomics **8**(3): 389-399.

Morris, T. J., L. M. Butcher, A. Feber, A. E. Teschendorff, A. R. Chakravarthy, T. K. Wojdacz and S. Beck (2014). "ChAMP: 450k Chip Analysis Methylation Pipeline." Bioinformatics **30**(3): 428-430.

Munoz-Fontela, C., M. A. Garcia, M. Collado, L. Marcos-Villar, P. Gallego, M. Esteban and C. Rivas (2007). "Control of virus infection by tumour suppressors." Carcinogenesis **28**(6): 1140-1144.

Murphy, A. K., M. Fitzgerald, T. Ro, J. H. Kim, A. I. Rabinowitsch, D. Chowdhury, C. L. Schildkraut and J. A. Borowiec (2014). "Phosphorylated RPA recruits PALB2 to stalled DNA replication forks to facilitate fork recovery." J Cell Biol **206**(4): 493-507.

Newell-Price, J., A. J. Clark and P. King (2000). "DNA methylation and silencing of gene expression." Trends Endocrinol Metab **11**(4): 142-148.

Newman, A. M., C. L. Liu and M. R. Green (2015). "Robust enumeration of cell subsets from tissue expression profiles." **12**(5): 453-457.

Nijwening, J. H., E. J. Geutjes, R. Bernards and R. L. Beijersbergen (2011). "The histone demethylase Jarid1b (Kdm5b) is a novel component of the Rb pathway and associates with E2f-target genes in MEFs during senescence." PLoS One **6**(9): e25235.

Ojesina, A. I., L. Lichtenstein, S. S. Freeman, C. S. Pedamallu, I. Imaz-Rosshandler, T. J. Pugh, A. D. Cherniack, L. Ambrogio, K. Cibulskis, B. Bertelsen, S. Romero-Cordoba, V. Trevino, K. Vazquez-Santillan, A. S. Guadarrama, A. A. Wright, M. W. Rosenberg, F. Duke, B. Kaplan, R. Wang, E. Nickerson, H. M. Walline, M. S. Lawrence, C. Stewart, S. L. Carter, A. McKenna, I. P. Rodriguez-Sanchez, M. Espinosa-Castilla, K. Woie, L. Bjorge, E. Wik, M. K. Halle, E. A. Hoivik, C. Krakstad, N. B. Gabino, G. S. Gomez-Macias, L. D. Valdez-Chapa, M. L. Garza-Rodriguez, G. Maytorena, J. Vazquez, C. Rodea, A. Cravioto, M. L. Cortes, H. Greulich, C. P. Crum, D. S. Neuberg, A. Hidalgo-Miranda, C. R. Escareno, L. A. Akslen, T. E. Carey, O. K. Vintermyr, S. B. Gabriel, H. A. Barrera-Saldana, J. Melendez-Zajgla, G. Getz, H. B. Salvesen and M. Meyerson (2014). "Landscape of genomic alterations in cervical carcinomas." Nature **506**(7488): 371-375.

Omberg, L., K. Ellrott, Y. Yuan, C. Kandoth, C. Wong, M. R. Kellen, S. H. Friend, J. Stuart, H. Liang and A. A. Margolin (2013). "Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas." Nat Genet **45**(10): 1121-1126.

Orlando, P. A., J. S. Brown, R. A. Gatenby and A. R. Guliano (2013). "The ecology of human papillomavirus-induced epithelial lesions and the role of somatic evolution in their progression." J Infect Dis **208**(3): 394-402.

Ottensmeier, C. H., K. L. Perry, E. L. Harden, J. Stasakova, V. Jenei, J. Fleming, O. Wood, J. Woo, C. H. Woelk, G. J. Thomas and S. M. Thirdborough (2016). "Upregulated glucose metabolism correlates inversely with CD8+ T cell infiltration and survival in squamous cell carcinoma." Cancer Res.

Paavonen, J., P. Naud, J. Salmeron, C. M. Wheeler, S. N. Chow, D. Apter, H. Kitchener, X. Castellsague, J. C. Teixeira, S. R. Skinner, J. Hedrick, U. Jaisamrarn, G. Limson, S. Garland, A. Szarewski, B. Romanowski, F. Y. Aoki, T. F. Schwarz, W. A. Poppe, F. X. Bosch, D. Jenkins, K. Hardt, T. Zahaf, D. Descamps, F. Struyf, M. Lehtinen and G. Dubin (2009). "Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by oncogenic HPV types (PATRICIA): final analysis of a double-blind, randomised study in young women." Lancet **374**(9686): 301-314.

Palefsky, J. (2009). "Human papillomavirus-related disease in people with HIV." Curr Opin HIV AIDS **4**(1): 52-56.

Pardoll, D. M. (2012). "The blockade of immune checkpoints in cancer immunotherapy." Nat Rev Cancer **12**(4): 252-264.

Parfenov, M., C. S. Pedamallu, N. Gehlenborg, S. S. Freeman, L. Danilova, C. A. Bristow, S. Lee, A. G. Hadjipanayis, E. V. Ivanova, M. D. Wilkerson, A. Protopopov, L. Yang, S. Seth, X. Song, J. Tang, X. Ren, J. Zhang, A. Pantazi, N. Santoso, A. W. Xu, H. Mahadeshwar, D. A. Wheeler, R. I. Haddad, J. Jung, A. I. Ojesina, N. Issaeva, W. G. Yarbrough, D. N. Hayes, J. R. Grandis, A. K. El-Naggar, M. Meyerson, P. J. Park, L. Chin, J. G. Seidman, P. S. Hammerman and R. Kucherlapati (2014). "Characterization of HPV and host genome interactions in primary head and neck cancers." Proc Natl Acad Sci U S A.

Park, J. and J. E. Schwarzbauer (2014). "Mammary epithelial cell interactions with fibronectin stimulate epithelial-mesenchymal transition." Oncogene **33**(13): 1649-1657.

Park, J. S., E. J. Kim, H. J. Kwon, E. S. Hwang, S. E. Namkoong and S. J. Um (2000). "Inactivation of interferon regulatory factor-1 tumor suppressor protein by HPV E7 oncoprotein. Implication for the E7-mediated immune evasion mechanism in cervical carcinogenesis." J Biol Chem **275**(10): 6764-6769.

Park, J. W., M. K. Shin, H. C. Pitot and P. F. Lambert (2013). "High Incidence of HPV-Associated Head and Neck Cancers in FA Deficient Mice Is Associated with E7's Induction of DNA Damage through Its Inactivation of Pocket Proteins." PLoS One **8**(9): e75056.

Peng, H., M. Talebzadeh-Farrooji, M. J. Osborne, J. W. Prokop, P. C. McDonald, J. Karar, Z. Hou, M. He, E. Kebebew, T. Orntoft, M. Herlyn, A. J. Caton, W. Fredericks, B. Malkowicz, C. S. Paterno, A. S. Carolin, D. W. Speicher, E. Skordalakes, Q. Huang, S. Dedhar, K. L. Borden and F. J. Rauscher, 3rd (2014). "LIMD2 is a small LIM-only protein overexpressed in metastatic lesions that regulates cell motility and tumor progression by directly binding to and activating the integrin-linked kinase." Cancer Res **74**(5): 1390-1403.

Perez-Ordonez, B., M. Beauchemin and R. C. Jordan (2006). "Molecular biology of squamous cell carcinoma of the head and neck." J Clin Pathol **59**(5): 445-453.

Pett, M. and N. Coleman (2007). "Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis?" J Pathol **212**(4): 356-367.

Pyeon, D., M. A. Newton, P. F. Lambert, J. A. den Boon, S. Sengupta, C. J. Marsit, C. D. Woodworth, J. P. Connor, T. H. Haugen, E. M. Smith, K. T. Kelsey, L. P. Turek and P. Ahlquist (2007). "Fundamental Differences in Cell Cycle Deregulation in Human Papillomavirus–Positive and Human Papillomavirus–Negative Head/Neck and Cervical Cancers." Cancer Research **67**(10): 4605-4619.

Rampias, T., C. Sasaki, P. Weinberger and A. Psyrri (2009). "E6 and e7 gene silencing and transformed phenotype of human papillomavirus 16-positive oropharyngeal cancer cells." J Natl Cancer Inst **101**(6): 412-423.

Reinius, L. E., N. Acevedo, M. Joerink, G. Pershagen, S. E. Dahlen, D. Greco, C. Soderhall, A. Scheynius and J. Kere (2012). "Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility." PLoS One **7**(7): e41361.

Reymond, N., B. B. d'Agua and A. J. Ridley (2013). "Crossing the endothelial barrier during metastasis." Nat Rev Cancer **13**(12): 858-870.

Rizvi, N. A., M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher and T. A. Chan (2015). "Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer." Science **348**(6230): 124-128.

Roberts, S. A. and D. A. Gordenin (2014). "Clustered and genome-wide transient mutagenesis in human cancers: Hypermutation without permanent mutators or loss of fitness." BioEssays : news and reviews in molecular, cellular and developmental biology.

Roberts, S. A., M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena, S. Harris, R. R. Shah, M. A. Resnick, G. Getz and D. A. Gordenin (2013). "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers." Nat Genet **45**(9): 970-976.

Roberts, S. A., M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena, S. Harris, R. R. Shah, M. A. Resnick, G. Getz and D. A. Gordenin (2013). "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers." Nat Genet **45**(9): 970-976.

Ronco, L. V., A. Y. Karpova, M. Vidal and P. M. Howley (1998). "Human papillomavirus 16 E6 oncoprotein binds to interferon regulatory factor-3 and inhibits its transcriptional activity." Genes Dev **12**(13): 2061-2072.

Rosenthal, R., N. McGranahan, J. Herrero, B. S. Taylor and C. Swanton (2016). "DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution." Genome Biol **17**: 31.

Roulois, D., H. Loo Yau, R. Singhania, Y. Wang, A. Danesh, S. Y. Shen, H. Han, G. Liang, P. A. Jones, T. J. Pugh, C. O'Brien and D. D. De Carvalho (2015). "DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts." Cell **162**(5): 961-973.

Sakofsky, C. J., S. A. Roberts, E. Malc, P. A. Mieczkowski, M. A. Resnick, D. A. Gordenin and A. Malkova (2014). "Break-induced replication is a source of mutation clusters underlying kataegis." Cell Rep **7**(5): 1640-1648.

Sanborn, A. L., S. S. Rao, S. C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K.

Stamenova, E. S. Lander and E. L. Aiden (2015). "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes." Proc Natl Acad Sci U S A **112**(47): E6456-6465.

Santegoets, L. A., M. Seters, T. J. Helmerhorst, C. Heijmans-Antonissen, P. Hanifi-Moghaddam, P. C. Ewing, W. F. van Ijcken, P. J. van der Spek, W. I. van der Meijden and L. J. Blok (2007). "HPV related VIN: highly proliferative and diminished responsiveness to extracellular signals." Int J Cancer **121**(4): 759-766.

Saraiya, M., E. R. Unger, T. D. Thompson, C. F. Lynch, B. Y. Hernandez, C. W. Lyu, M. Steinau, M. Watson, E. J. Wilkinson, C. Hopenhayn, G. Copeland, W. Cozen, E. S. Peters, Y. Huang, M. S. Saber, S. Altekruse and M. T. Goodman (2015). "US assessment of HPV types in cancers: implications for current and 9-valent HPV vaccines." J Natl Cancer Inst **107**(6): djv086.

Schiffman, M., P. E. Castle, J. Jeronimo, A. C. Rodriguez and S. Wacholder (2007). "Human papillomavirus and cervical cancer." Lancet **370**(9590): 890-907.

Scotto, L., G. Narayan, S. V. Nandula, H. Arias-Pulido, S. Subramaniyam, A. Schneider, A. M. Kaufmann, J. D. Wright, B. Pothuri, M. Mansukhani and V. V. Murty (2008). "Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression." Genes Chromosomes Cancer **47**(9): 755-765.

Shi, W., Z. Feng, J. Zhang, I. Gonzalez-Suarez, R. P. Vanderwaal, X. Wu, S. N. Powell, J. L. Roti Roti, S. Gonzalo and J. Zhang (2010). "The role of RPA2 phosphorylation in homologous recombination in response to replication arrest." Carcinogenesis **31**(6): 994-1002.

Skobe, M., T. Hawighorst, D. G. Jackson, R. Prevo, L. Janes, P. Velasco, L. Riccardi, K. Alitalo, K. Claffey and M. Detmar (2001). "Induction of tumor lymphangiogenesis by VEGF-C promotes breast cancer metastasis." Nat Med **7**(2): 192-198.

Slebos, R. J., Y. Yi, K. Ely, J. Carter, A. Evjen, X. Zhang, Y. Shyr, B. M. Murphy, A. J. Cmelak, B. B. Burkey, J. L. Netterville, S. Levy, W. G. Yarbrough and C. H. Chung (2006). "Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma." Clin Cancer Res **12**(3 Pt 1): 701-709.

Smogorzewska, A., S. Matsuoka, P. Vinciguerra, E. R. McDonald Iii, K. E. Hurov, J. Luo, B. A. Ballif, S. P. Gygi, K. Hofmann, A. D. D'Andrea and S. J. Elledge (2007). "Identification of the FANCI Protein, a Monoubiquitinated FANCD2 Paralog Required for DNA Repair." Cell **129**(2): 289-301.

Snyder, A., V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok and T. A. Chan (2014). "Genetic basis for clinical response to CTLA-4 blockade in melanoma." N Engl J Med **371**(23): 2189-2199.

Spardy, N., A. Duensing, D. Charles, N. Haines, T. Nakahara, P. F. Lambert and S. Duensing (2007). "The human papillomavirus type 16 E7 oncoprotein activates the Fanconi anemia (FA) pathway and causes accelerated chromosomal instability in FA cells." J Virol **81**(23): 13265-13270.

Stanley, M. A. (2012). "Epithelial cell responses to infection with human papillomavirus." Clin Microbiol Rev **25**(2): 215-222.

Stevanovic, S., L. M. Draper, M. M. Langhan, T. E. Campbell, M. L. Kwong, J. R. Wunderlich, M. E. Dudley, J. C. Yang, R. M. Sherry, U. S. Kammula, N. P. Restifo, S. A. Rosenberg and C. S. Hinrichs (2015). "Complete regression of metastatic cervical cancer after treatment with human papillomavirus-targeted tumor-infiltrating T cells." J Clin Oncol **33**(14): 1543-1550.

Stransky, N., A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis, A. Sivachenko, G. V. Kryukov, M. S. Lawrence, C. Sougnez, A. McKenna, E. Shefler, A. H. Ramos, P. Stojanov, S. L. Carter, D. Voet,

M. L. Cortes, D. Auclair, M. F. Berger, G. Saksena, C. Guiducci, R. C. Onofrio, M. Parkin, M. Romkes, J. L. Weissfeld, R. R. Seethala, L. Wang, C. Rangel-Escareno, J. C. Fernandez-Lopez, A. Hidalgo-Miranda, J. Melendez-Zajgla, W. Winckler, K. Ardlie, S. B. Gabriel, M. Meyerson, E. S. Lander, G. Getz, T. R. Golub, L. A. Garraway and J. R. Grandis (2011). "The mutational landscape of head and neck squamous cell carcinoma." Science **333**(6046): 1157-1160.

Stricker, S. H., A. Feber, P. G. Engstrom, H. Caren, K. M. Kurian, Y. Takashima, C. Watts, M. Way, P. Dirks, P. Bertone, A. Smith, S. Beck and S. M. Pollard (2013). "Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner." Genes Dev **27**(6): 654-669.

Sturm, D., B. A. Orr, U. H. Toprak, V. Hovestadt, D. T. Jones, D. Capper, M. Sill, I. Buchhalter, P. A. Northcott, I. Leis, M. Ryzhova, C. Koelsche, E. Pfaff, S. J. Allen, G. Balasubramanian, B. C. Worst, K. W. Pajtler, S. Brabetz, P. D. Johann, F. Sahm, J. Reimand, A. Mackay, D. M. Carvalho, M. Remke, J. J. Phillips, A. Perry, C. Cowdrey, R. Drissi, M. Fouladi, F. Giangaspero, M. Lastowska, W. Grajkowska, W. Scheurlen, T. Pietsch, C. Hagel, J. Gojo, D. Lotsch, W. Berger, I. Slavc, C. Haberler, A. Jouvet, S. Holm, S. Hofer, M. Prinz, C. Keohane, I. Fried, C. Mawrin, D. Scheie, B. C. Mobley, M. J. Schniederjan, M. Santi, A. M. Buccoliero, S. Dahiya, C. M. Kramm, A. O. von Bueren, K. von Hoff, S. Rutkowski, C. Herold-Mende, M. C. Fruhwald, T. Milde, M. Hasselblatt, P. Wesseling, J. Rossler, U. Schuller, M. Ebinger, J. Schittenhelm, S. Frank, R. Grobholz, I. Vajtai, V. Hans, R. Schneppenheim, K. Zitterbart, V. P. Collins, E. Aronica, P. Varlet, S. Puget, C. Dufour, J. Grill, D. Figarella-Branger, M. Wolter, M. U. Schuhmann, T. Shalaby, M. Grotzer, T. van Meter, C. M. Monoranu, J. Felsberg, G. Reifenberger, M. Snuderl, L. A. Forrester, J. Koster, R. Versteeg, R. Volckmann, P. van Sluis, S. Wolf, T. Mikkelsen, A. Gajjar, K. Aldape, A. S. Moore, M. D. Taylor, C. Jones, N. Jabado, M. A. Karajannis, R. Eils, M. Schlesner, P. Lichter, A. von Deimling, S. M. Pfister, D. W. Ellison, A. Korshunov and M. Kool (2016). "New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs." Cell **164**(5): 1060-1072.

Suzuki, R. and H. Shimodaira (2006). "Pvclust: an R package for assessing the uncertainty in hierarchical clustering." Bioinformatics **22**(12): 1540-1542.

Tang, K.-W., B. Alaei-Mahabadi, T. Samuelsson, M. Lindh and E. Larsson (2013). "The landscape of viral expression and host gene fusion and adaptation in human cancer." Nat Commun **4**.

Tang, S., M. Tao, J. P. McCoy, Jr. and Z. M. Zheng (2006). "The E7 oncoprotein is translated from spliced E6*I transcripts in high-risk human papillomavirus type 16- or type 18-positive cervical cancer cell lines via translation reinitiation." J Virol **80**(9): 4249-4263.

TCGA (2015). "Comprehensive genomic characterization of head and neck squamous cell carcinomas." Nature **517**(7536): 576-582.

Teschendorff, A. E., F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero and S. Beck (2013). "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data." Bioinformatics **29**(2): 189-196.

Timp, W., H. C. Bravo, O. G. McDonald, M. Goggins, C. Umbricht, M. Zeiger, A. P. Feinberg and R. A. Irizarry (2014). "Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors." Genome Med **6**(8): 61.

Triche, T. J., Jr., D. J. Weisenberger, D. Van Den Berg, P. W. Laird and K. D. Siegmund (2013). "Low-level processing of Illumina Infinium DNA Methylation BeadArrays." Nucleic Acids Res **41**(7): e90.

Tsai, H. C., H. Li, L. Van Neste, Y. Cai, C. Robert, F. V. Rassool, J. J. Shin, K. M. Harbom, R. Beaty, E. Pappou, J. Harris, R. W. Yen, N. Ahuja, M. V. Brock, V. Stearns, D. Feller-Kopman, L. B. Yarmus, Y. C. Lin, A. L. Welm, J. P. Issa, I. Minn, W. Matsui, Y. Y. Jang, S. J. Sharkis, S. B. Baylin and C. A.

Zahnow (2012). "Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells." Cancer Cell **21**(3): 430-446.

Unsal-Kacmaz, K., T. E. Mullen, W. K. Kaufmann and A. Sancar (2005). "Coupling of Human Circadian and Cell Cycles by the Timeless Protein." Mol Cell Biol **25**(8): 3109-3116.

van Houten, V. M., P. J. Snijders, M. W. van den Brekel, J. A. Kummer, C. J. Meijer, B. van Leeuwen, F. Denkers, L. E. Smeele, G. B. Snow and R. H. Brakenhoff (2001). "Biological evidence that human papillomaviruses are etiologically involved in a subgroup of head and neck squamous cell carcinomas." Int J Cancer **93**(2): 232-235.

Vartanian, J.-P., D. Guetard, M. Henry and S. Wain-Hobson (2008). "Evidence for Editing of Human Papillomavirus DNA by APOBEC3 in Benign and Precancerous Lesions." Science **320**(5873): 230-233.

Vermeulen, K., D. R. Van Bockstaele and Z. N. Berneman (2003). "The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer." Cell Prolif **36**(3): 131-149.

Vokes, E. E., N. Agrawal and T. Y. Seiwert (2015). "HPV-Associated Head and Neck Cancer." Journal of the National Cancer Institute **107**(12).

von Eyss, B., J. Maaskola, S. Memczak, K. Mollmann, A. Schuetz, C. Loddenkemper, M. D. Tanh, A. Otto, K. Muegge, U. Heinemann, N. Rajewsky and U. Ziebold (2012). "The SNF2-like helicase HELLS mediates E2F3-dependent transcription and cellular transformation." Embo j **31**(4): 972-985.

Wang, H., P. Mo, S. Ren and C. Yan (2010). "Activating transcription factor 3 activates p53 by preventing E6-associated protein from binding to E6." J Biol Chem **285**(17): 13201-13210.

Ward, M. J., S. M. Thirdborough, T. Mellows, C. Riley, S. Harris, K. Suchak, A. Webb, C. Hampton, N. N. Patel, C. J. Randall, H. J. Cox, S. Jogai, J. Primrose, K. Piper, C. H. Ottensmeier, E. V. King and G. J. Thomas (2014). "Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer." Br J Cancer **110**(2): 489-500.

Weaving, L. S., C. J. Ellaway, J. Gecz and J. Christodoulou (2005). "Rett syndrome: clinical review and genetic update." J Med Genet **42**(1): 1-7.

Wei, Q. (2005). "Pitx2a binds to human papillomavirus type 18 E6 protein and inhibits E6-mediated P53 degradation in HeLa cells." J Biol Chem **280**(45): 37790-37797.

Westra, W. H., J. M. Taube, M. L. Poeta, S. Begum, D. Sidransky and W. M. Koch (2008). "Inverse relationship between human papillomavirus-16 infection and disruptive p53 gene mutations in squamous cell carcinoma of the head and neck." Clin Cancer Res **14**(2): 366-369.

Wu, J., J. Li, R. Salcedo, N. F. Mivechi, G. Trinchieri and A. Horuzsko (2012). "The proinflammatory myeloid cell receptor TREM-1 controls Kupffer cell activation and development of hepatocellular carcinoma." Cancer Res **72**(16): 3977-3986.

Yan, J., Q. Li, S. Lievens, J. Tavernier and J. You (2010). "Abrogation of the Brd4-positive transcription elongation factor B complex by papillomavirus E2 protein contributes to viral oncogene repression." J Virol **84**(1): 76-87.

Yang, X., H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones and G. Liang (2014). "Gene body methylation can alter gene expression and is a therapeutic target in cancer." Cancer Cell **26**(4): 577-590.

Yeh, E., M. Cunningham, H. Arnold, D. Chasse, T. Monteith, G. Ivaldi, W. C. Hahn, P. T. Stukenberg, S. Shenolikar, T. Uchida, C. M. Counter, J. R. Nevins, A. R. Means and R. Sears (2004). "A signalling pathway controlling c-Myc degradation that impacts oncogenic transformation of human cells." Nat Cell Biol **6**(4): 308-318.

Yoshihara, K., M. Shahmoradgoli, E. Martinez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Trevino, H. Shen, P. W. Laird, D. A. Levine, S. L. Carter, G. Getz, K. Stemke-Hale, G. B. Mills and R. G. Verhaak (2013). "Inferring tumour purity and stromal and immune cell admixture from expression data." Nat Commun **4**: 2612.

Yuan, Z., H. J. Mehta, K. Mohammed, N. Nasreen, R. Roman, M. Brantly and R. T. Sadikot (2014). "TREM-1 is induced in tumor associated macrophages by cyclo-oxygenase pathway in human non-small cell lung cancer." PLoS One **9**(5): e94241.

Zhai, Y., R. Kuick, B. Nan, I. Ota, S. J. Weiss, C. L. Trimble, E. R. Fearon and K. R. Cho (2007). "Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion." Cancer Res **67**(21): 10163-10172.

Zhang, Y., B. Liu, Q. Zhao, T. Hou and X. Huang (2014). "Nuclear localizaiton of beta-catenin is associated with poor survival and chemo-/radioresistance in human cervical squamous cell cancer." Int J Clin Exp Pathol **7**(7): 3908-3917.

Zheng, Z. M. and C. C. Baker (2006). "Papillomavirus genome structure, expression, and post-transcriptional regulation." Front Biosci **11**: 2286-2302.

Appendices

A1. List of IPA canonical pathways for the HPV-associated gene expression metasignature.

Ingenuity Canonical Pathways	-log(B-H p-value)	Ratio	z-score
Cell Cycle Control of Chromosomal Replication	1.54E+01	3.42E-01	NaN
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	9.23E+00	2.04E-01	-2.121
Mitotic Roles of Polo-Like Kinase	5.55E+00	1.27E-01	1.633
Cyclins and Cell Cycle Regulation	4.98E+00	1.04E-01	0.378
GADD45 Signaling	4.84E+00	2.63E-01	NaN
Mismatch Repair in Eukaryotes	3.62E+00	2.50E-01	NaN
DNA damage-induced 14-3-3ϊ*f* Signaling	3.37E+00	2.11E-01	NaN
p53 Signaling	3.06E+00	6.31E-02	-0.447
Estrogen-mediated S-phase Entry	3.05E+00	1.67E-01	1
Granulocyte Adhesion and Diapedesis	2.90E+00	4.85E-02	NaN
Cell Cycle: G1/S Checkpoint Regulation	2.53E+00	7.94E-02	1
Role of BRCA1 in DNA Damage Response	2.14E+00	6.41E-02	NaN
ATM Signaling	2.14E+00	6.33E-02	-0.447
Agranulocyte Adhesion and Diapedesis	2.08E+00	4.00E-02	NaN

dTMP De Novo Biosynthesis	2.05E+00	4.00E-01	NaN
Hereditary Breast Cancer Signaling	1.91E+00	4.32E-02	NaN
Role of CHK Proteins in Cell Cycle Checkpoint Control	1.90E+00	7.27E-02	NaN
BER pathway	1.33E+00	1.67E-01	NaN
Small Cell Lung Cancer Signaling	1.30E+00	4.76E-02	NaN

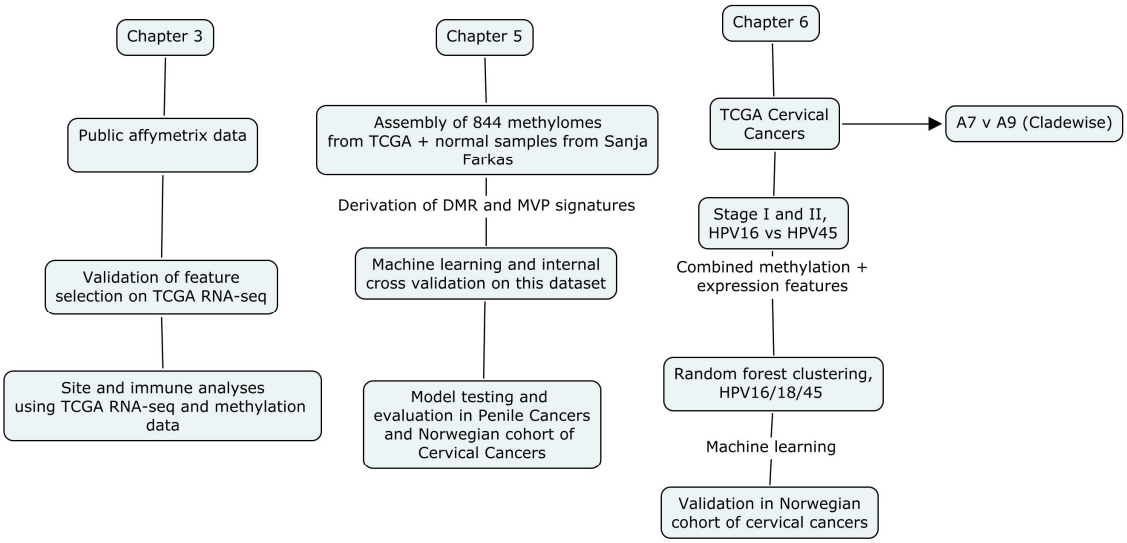
A2: Upstream Regulatory Analysis of HPV-associated metasignature.

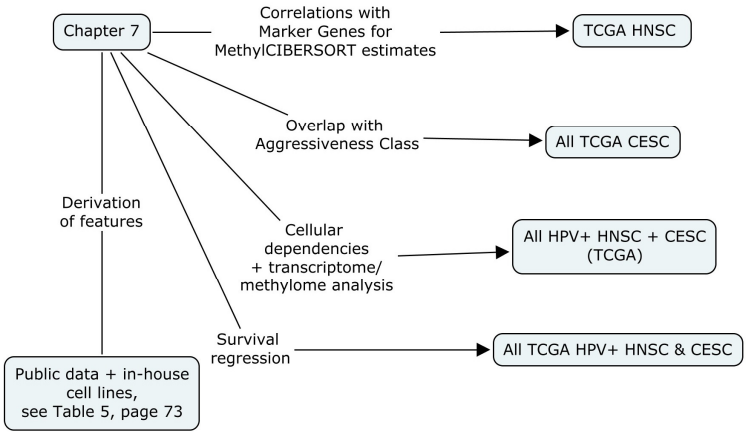
Upstream Regulator	Exp Fold Change	Molecule Type	Predicted Activation State	Activation z-score	Flags	p-value of overlap
NUPR1		transcription regulator	Inhibited	-3.638		5.24E-08
CDKN2A	19.638	transcription regulator	Inhibited	-2.205	bias	3.24E-05
ATF3		transcription regulator	Inhibited	-2		1.45E-05
E2F6		transcription regulator	Inhibited	-2.646	bias	2.64E-12
KDM5B		transcription regulator	Inhibited	-3.966	bias	6.91E-17
TP53		transcription regulator	Inhibited	-5.45		1.18E-28
CREB1		transcription regulator	Activated	2.646	bias	1.03E-08
FOXM1	2.439	transcription regulator	Activated	3.618	bias	9.43E-26
FOXO1		transcription regulator	Activated	3.688	bias	3.99E-17
E2F3		transcription regulator	Activated	2.53		4.97E-09
MITF		transcription regulator	Activated	3.873	bias	1.12E-16
TAL1		transcription regulator	Activated	2.496		3.38E-10
RARA		ligand-dependent nuclear receptor	Activated	3		1.62E-05

<continued on the next page>

A3: Analysis Workflows

These flowcharts represent the bulk of the analyses that involved multiple datasets in generating the results found in Chapters 3, 5,6 and 7.





Appendix A4 : Clinical variables in the TCGA HPV+ HNSC cohort and survival analysis coefficients from Cox regression. Table reproduced from (Chakravarthy, Henderson et al. 2016)

				Multivariable Analysis‡ (OS)	
Variable	OPSCC	*P*†	Non-OPSCC	HR (95% CI)	*P*
Median age (range), years	56 (35-77)	.0066	59.5 (49–82)	0.95 (0.88 to 1.02)	.47
Smoking, pack-years, No.				1.72 (0.4 to 7.69)	.47
< 10	27		5		
> 10	20		9		
		.23			
T stage, No.				0.76 (0.16 to 3.57)	.72
1/2	36		7		
3/4	15		11		
		.03			
N stage, No.				0.47 (0.12 to 1.85)	.276
N0-N2a	27		12		
N2b-N3	26		5		
		.17			
Immune cluster, No.				0.16 (0.03 to 0.85)	.031
Depleted	19		14		
Enriched	35		4		
		.002			

Abbreviations: HR, hazard ratio; OPSCC, oropharyngeal squamous cell carcinoma; OS, overall survival.
*Comparisons between OPSCC and non-OPSCC were conducted using all samples in the dataset for which the relevant metadata were available.
†*P* values were generated using Wilcoxon test (age) and Fisher's exact test (smoking, stage, immune cluster).
‡Multivariable survival analysis was conducted using the subset of 58 samples for which data on all variables were available.

Appendix A5: K-M curves of survival by histology between HPV16-like and HPV45-like tumours in the Norwegian cohort of cervical cancers. The cohort only included 2 neuroendocrine cancers and ergo have not been plotted.